The Effect of Feature Hierarchies on Frequencies of Passivization in English

Shipra Dingare
August 2001

[1] This work is based in part on research supported by the National Science Foundation under Grant No. BCS-9818077 and by a gift to the Stanford Symbolic Systems Program by Ric Weiland.

Acknowledgments

It is my pleasure to thank Joan Bresnan, my principal advisor, for making this thesis seem manageable, and for pointing me in directions for research, and for consistently being enthusiastic and suggesting solutions to problem after problem, and equally, for suggesting problems and keeping me aware of the issues involved. And particularly for always making me feel more comfortable, in classes and meetings and a number of other situations, to express myself freely, and for handling it gracefully when I was incoherent. I was lucky to be able to work with her and it was a unique experience watching her work.

Thanks to Chris Manning, my second advisor, for being helpful in a number of ways despite the incredible demands on his time this year, particularly in providing advice with statistics, and helping me to search corpora, and setting up software for searching them, and for reading what I wrote critically, and in general for being down-to-earth.

Thanks to both Chris and Joan for giving me the opportunity to work on such an interesting project. I have really enjoyed our meetings and watching the ideas in the project take shape. Thanks especially for a wonderful trip to Hong Kong. And thanks to Tom Wasow for helping me get involved with the project and for teaching, with Ivan Sag, an introductory syntax class which I very much liked.

A very special thanks to Ivan Sag, who is almost entirely responsible, along with Tom Wasow, for my having had the opportunity to enter the coterminal master's program in linguistics. It was very kind of him to help me, on the basis of having known me a very short time, to pursue my education in linguistics further. Last year I enjoyed the classes I took with him and the research project I pursued at his suggestion for his class, and this year he has been consistently helpful, in helping me meet requirements for my degree, and in searching for jobs, and by pulling strings here and there when I seemed to have boxed myself into a bureaucratic corner.

Finally, I would like to thank my mother, father, and sister for always being so indulgent of me, particularly my mother for making me feel welcome to wake her up at strange hours so I could complain. And I can't thank my aunt, uncle, and cousin, who I lived with this year, enough for how kind they've always been to me, particularly in never complaining that I monopolized their computer, and for very considerately always keeping my favorite cereal in the cabinet and lots of soda in the fridge.

CONTENTS

Chapter 1

Introduction

Prominence hierarchies along various dimensions have been posited to play a role in various syntactic phenomena in diverse languages, particularly within typological approaches to linguistics. These hierarchies have been grounded in different ways by different researchers, including: the tendency for elements higher on the hierarchy to be topics, to be agents, to be more mentally accessible, to be easier for the speaker to empathize with, and so on. In all of these approaches, however, there is every reason to believe that their effects should be felt not only in the formal properties of a few particular languages, but in every language. In what follows I examine two particular hierarchies - the hierarchy of person and the hierarchy of definiteness – and explore their influence not on the grammaticality but on the *frequency* of passivization in English. Demonstrating that these hierarchies have an effect on frequencies of passivization supports and quantifies the speculation that the hierarchies are rooted in psycholinguistic or communicative tendencies. At the same time, the frequency results can be taken to support the idea that frequencies and gradations in frequency are *principled*. This will be a secondary goal of the present work. These effects will be formalized and modeled using the harmonic alignment technique described in Aissen (1999, 2000) and the Stochastic OT framework of Boersma and Hayes (2001). Finally, in the conclusion I explore hierarchy overlap and how the stochastic OT learning algorithm restricts the typology of possible languages beyond what is predicted by standard factorial typologies.

1.1 Hierarchies

The evidence for positing a hierarchy generally consists in presenting parallel phenomena in various languages and demonstrating that they make reference to different cut-off points on the hierarchy. For example, Aissen (2000) motivates the ordering “Pronoun > Proper Noun > Definite > Indefinite Specific > Non-Specific” for the definiteness hierarchy by pointing out that in the phenomenon of differential object marking, there are languages which mark only pronoun objects, or only pronoun and proper noun objects, or pronoun , proper noun, and definite objects, but no languages which mark definite objects but do not mark pronoun or proper noun objects. Similarly, Silverstein (1976) motivates the ordering "Pronoun > Proper Noun > Human > Animate > Inanimate” by arguing that in the phenomenon of split ergativity, if a language forces ergative case-marking on agents at some point in the hierarchy, then it also marks agents at all points below that point in the hierarchy, and that if a language marks patients with accusative case-marking at some point in the hierarchy, then it will also mark patients above that point in the hierarchy.

An explanation of a hierarchy must go further than simply providing evidence that languages make reference to a particular hierarchy in particular phenomena. To explain why a hierarchy is influential cross-linguistically, it must be shown that it is “rooted” in some way that causes it to have effects in more than one language. For example, elements at the high end of hierarchies have been associated with greater cognitive salience, higher frequency of topicality, higher likelihood of being agents, and so on. A theory of hierarchies must connect this grounding of a hierarchy with the phenomena to which it will be relevant; that is, it must predict which phenomena will make reference to a particular hierarchy. In general, researchers attempting to “ground” a particular hierarchy do so in different ways depending on the phenomenon they are trying to explain, and their approach often does not explain other phenomena which make reference to the same hierarchy. In the chapters to follow, I will review for each hierarchy the different ways in which researchers have attempted to explain its influence, and which of those theories implies that the hierarchy will influence the choice between active and passive. Furthermore, evidence will be provided that these hierarchies *do* influence this choice statistically in English, thereby supporting those explanations that predict an interaction with passivization, and supporting the need for a theory that predicts effects of hierarchies not only in a few isolated languages, but in potentially all languages.

1.2 Frequency

Any study that uses frequency data to study phenomena related to grammar must justify doing so. Due to the long-held distinction between competence and performance, matters of frequency have long been outside of what was deemed worthwhile to be studied by syntacticians. Nevertheless, in the search for cross-linguistic universals it has sporadically been noted that grammatical phenomena in certain languages are mirrored by frequentistic phenomena in others, supporting the idea that frequencies are principled in the same way that grammars are. This point is made forcefully in Givón (1979) in a passage challenging the competence-performance distinction:

> In many of the world's languages, probably in most, the subject of declarative clauses cannot be referential-indefinite...Languages of this type are, for example, Swahili, Bemba, Rwanda (Bantu), Chinese, Sherpa (Sino-Tibetan), Bikol (Austronesian), Ute (Uto-Aztecan), Krio (Creole), all Creoles, and many others…In a relatively small number of the world's languages…referential-indefinite nouns may appear as subjects of nonpresentative sentences…When one investigates the text frequency of [such] sentences in English, however, one finds them at an extremely low frequency: About 10% of the subjects of main-declarative-affirmative-active sentences (nonpresentative) are indefinite, as against 90% definite. Now this is presumably not a fact about the "competence" of English speakers, but only about their actual "language behavior." But are we dealing with two different kinds of facts in English and Krio? Hardly. What we are dealing with is apparently the very same *communicative tendency* – to reserve the subject position in the sentence for the *topic*, the old-information argument, the "continuity marker." In some languages, (Krio, etc.) this communicative tendency is expressed at the *categorial* level of 100%. In other languages (English, etc.) the very same communicative tendency is expressed "only" at the *noncategorial* level of 90%. And a transformational-generative linguist will then be forced to count this fact as competence in Krio and performance in English. But what is the communicative difference between a rule of 90% fidelity and one of 100% fidelity? In psychological terms, next to nothing…When live discourse data are taken into account…it becomes obvious that noncategorial phenomena are the **rule** rather than the exception in human language. (pp.26-31)

Givón goes on to provide additional data on phenomena including agentless passivization and indefinite objects under negation, further supporting the contention that phenomena which are categorical in some languages are statistical tendencies in others. The link between frequency and grammaticality is also made by Winter (1971), who shows for case marking that more frequent forms are more likely to survive than less frequent forms. Greenbaum (1980) shows that there is an association between acceptability judgments and perceived frequencies. That is, sentences that are perceived to be more frequent are more likely to be judged more acceptable; this point is also made in Boersma and Hayes (2001). Greenberg (1966) cites frequency as evidence for markedness - in arguing for the hierarchy "singular > plural > dual" cites frequency data. The close link between frequency and grammaticality supports the idea that principles known to influence grammars also influence frequencies. This, together with the fact that prominence hierarchies have been shown to drive categorical phenomena in various languages, makes it reasonable to expect effects of prominence hierarchies on frequencies in English.

Studying frequencies also allows us to "quantify" hierarchies. One of the questions that remains relatively unexplored, is the relation or "distance" between various elements on the hierarchies. For example, given the Silverstein hierarchy mentioned above, one might expect "cut-off points" to be chosen anywhere in the hierarchy. That is, there is no reason to expect some cut-off points to be more frequent than others. Similarly, in the definiteness hierarchy ("pronoun > proper noun > definite > indefinite specific > non-specific") there is no specification as to whether pronouns are ranked more closely to proper nouns, or proper nouns ranked more closely to definites, than definites to indefinites, for example. That such a ranking is necessary is supported by the fact that some cut-off points are more likely than others. Silverstein himself notes in regard to split-ergative languages that "simple, binary, two–way splits usually are defined around some feature F_i from among those of person"; DeLancey (1981) notes that splits centering on the Local Person > 3^{rd} distinction and the pronoun > full NP distinction are the most common, while all others are quite rare. Studying frequency effects of the hierarchies allows one to quantify the extent to which positions on the hierarchy differ in distance.

1.3 Stochastic Optimality Theory and The Gradual Learning Algorithm

The effect of linguistic constraints on frequency can be formalized in the Stochastic Optimality Theory approach of Boersma and Hayes (2000). Below, I first briefly review "vanilla" optimality theory and then go on to introduce the stochastic optimality theory framework.

1.3.1 "Vanilla" Optimality Theory

In standard optimality theory, a grammar is a function which provides for each input a structural description or output. While exactly what the "input" comprises varies from account to account, common assumptions in optimality-theoretic syntax are that the input consists of a predicate argument-structure specifying information such as tense and the semantic role or discourse prominence of each argument. It is assumed that universal grammar provides an infinite set of candidates for each input and a set of universal well-formedness constraints which provide the basis for choosing the optimal output for each input. While the constraints are universal, languages differ in how the constraints are *ranked*. Each language ranks the constraints from the highest-ranked to the lowest-ranked. It is hypothesized that all possible rankings of constraints represent all possible languages.

In the process of selecting an optimal candidate, each candidate is assessed for the number of times it violates each constraint. Then, all candidates are compared on the highest-ranked constraint. If any candidate has zero violations of this constraint, all candidates having one or more violations are eliminated. If all candidates have one or more violations, then one violation is subtracted from each candidate until there is a candidate having zero violations. At that point all candidates having one or more violations are eliminated. If there is more than one candidate left at that point, the candidates are compared with respect to the second constraint. This process is repeated until there is only one candidate left. This candidate is the "winner". An example from Prince and Smolensky (1997) is reviewed below. The constraints are NOCODA and PARSE and are listed with the highest-ranked constraint leftmost and the lowest-ranked rightmost. The input is at the top left and the candidates are listed below it. Asterisks represent constraint violations; "fatal" violations are marked with an exclamation mark. The hand points to the winner.

Tableau 1.

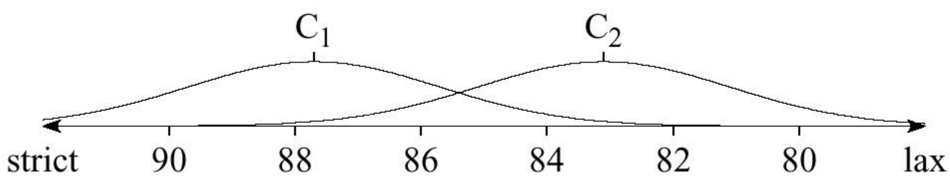
/batak/	NOCODA	PARSE
☞ [ba.ta]		*
*[ba]		**!*
*[ba.tak]	*!	
*[bat]	*!	**

In this example, the candidates are first evaluated on the highest-ranked constraint NOCODA. The candidates *ba.tak* and *bat* are eliminated since the candidates *ba.ta* and *ba* do not violate this constraint. Since there are two candidates left, these two candidates are evaluated on the next-highest ranked constraint, PARSE. Since no candidate has zero violations, one violation of PARSE is subtracted for *ba.ta* and *ba*. This leaves *ba.ta* with zero violations and *ba* with two violations, so *ba.ta* is the optimal candidate.

1.3.2 Stochastic Optimality Theory

Stochastic OT differs from standard OT in that it presupposes a continuous scale of constraint rankings rather than a discrete ordinal scale. That is, constraints do not simply have a rank – they have a real-number value. Also, Stochastic OT assumes that at every evaluation of a candidate set, a small amount of noise drawn from a normal distribution is added to the ranking value of each constraint. The real-number value permanently associated with the constraint (the value to which the noise is added) will be referred to as the *ranking value*. The constraint's value at the time of evaluation is referred to as the *selection point* (or the *ranking value at the time of evaluation*). The amount of noise is drawn from a normal distribution with a mean of zero and a fixed standard deviation (arbitrarily chosen as 2). Thus, a constraint's value is itself normally distributed, as depicted below, with the mean falling at the constraint's ranking value. The normal distribution is a probability density function representing the probability that a constraint will have a particular value or importance.

(1) *Ranking distributions for two constraints* (Boersma and Hayes 2001)



Because Stochastic OT presupposes a continuous scale of constraint rankings, constraints differ not only in dominance but in distance. The distance between two constraints is crucial to the predictions of the theory. If one constraint outranks another, then over a number of evaluations the higher-ranked constraint will outrank the other a majority of the time. However, depending on the distance between them, the lower-ranked constraint will outrank the higher-ranked constraint a certain percentage of the time. For example, in the picture above, constraint C_1 one will outrank constraint C_2 most of the time, but constraint C_2 will outrank constraint C_1 in a small percentage of evaluations.

Categorical rankings arise when two constraints are sufficiently far apart such that the odds of the lower constraint outranking the higher constraint become vanishingly low. For example, if two constraints are five standard deviations apart, the odds of the lower constraint outranking the higher constraint are approximately 1 in 5000. If two constraints are nine standard deviations apart, the odds of the lower constraint outranking the higher constraint are approximately 1 in 10 billion, meaning that this would probably not occur in a speaker's lifetime (Boersma & Hayes 2001).

Stochastic OT grammars are learned through the Gradual Learning Algorithm. The goal of algorithm is to learn the permanent ranking values associated with each constraint (the mean of the normal distribution). There is an initial state of constraint rankings which is whatever the linguist presumes the initial state of constraint rankings to be. In Boersma and Hayes (2001) it is assumed that all constraints start with a ranking value of 100. During training, a number of surface forms are presented. It is assumed that the algorithm is able to infer the input form from the surface form (cf. Tesar & Smolensky (1996) on "robust interpretive parsing"). Next, the algorithm takes the inferred input form and generates from the current grammar a surface form for the input form. To do this, it takes the current constraint rankings, perturbs each ranking with an amount of noise taken from a normal distribution, and uses the resulting rankings to evaluate a winner from the candidate set. If the winner matches the surface form provided, no adjustment is made. But if the form generated by the grammar does not match the surface form provided, the algorithm makes the following adjustment: the constraint violations of the correct candidate are compared to the constraint violations of the incorrect candidate chosen by the algorithm. All the constraint violations which the two candidates share are cancelled or ignored – this is called mark cancellation and is illustrated in (3). Then, all the constraints which the incorrect candidate violated are incremented by a certain amount called the plasticity. Also, all the constraints which the correct candidate violated are decremented by the same amount.

Example (from Boersma and Hayes 2000):

(2) Constraint Violations of Candidates 1 and 2

/underlying form/	C1	C2	C3	C4	C5	C6	C7	C8
√ Candidate 1 (learning datum)	*!	**	*		*			*
☞ Candidate 2 (learner's output)		*	*	*		*		*

(3) Mark Cancellation

/underlying form/	C1	C2	C3	C4	C5	C6	C7	C8
√ Candidate 1 (learning datum)	*!	*~~*~~	~~*~~		*			~~*~~
☞ Candidate 2 (learner's output)		~~*~~	~~*~~	*		*		~~*~~

(4) Marks Remaining After Mark Cancellation

/underlying form/	C1	C2	C3	C4	C5	C6	C7	C8
√ Candidate 1 (learning datum)	*	*			*			
☞ Candidate 2 (learner's output)				*		*		

Thus, constraints 1, 2, and 5 would be demoted, and constraints 4 and 6 would be promoted. The algorithm repeats these steps for training datum after training datum.

One also has the option of implementing subhierarchies of constraints. A subhierarchy refers to a set of constraints that must always be ranked in a specific order. For example, the subhierarchy $C_1 > C_2$ implies that while C_1 may be ranked anywhere, C_2 must be ranked below it. As will be seen, the optimality-theoretic formalization of prominence scales will result in such fixed subhierarchies of constraints. The fixed rankings are implemented in learning as follows: learning proceeds as usual; however, if at any point during learning, the ranking of C_1 is demoted so that it falls below C_2, the algorithm immediately demotes C_2 as well so that it remains ranked below C_1. Similarly if C_2 is promoted so that it is ranked above C_1. Note that it is only the permanent ranking values which are maintained in the fixed order; only the permanent ranking value of C_1 is guaranteed to be above C_2. At the time of evaluation, noise may cause C_2 to be ranked above C_1.

Once a constraint ranking is learned, one can sample with a large number of evaluations (all of these operations – training with the gradual learning algorithm, sampling – are implemented in the Praat system (Boersma and Weenink 2000)). As noted before, at each evaluation the constraint rankings are perturbed by a small amount of noise. Over a large number of evaluations, a frequency distribution of outputs for a particular input will appear. This frequency distribution can be compared with the frequency distribution in the training data. Clearly, with an arbitrary number and set of constraints one can model any frequency distribution and obtain a close match between the training data and the output data. Thus, a close match of the training data with the output data is not necessarily a sign of success. However, with a motivated set of constraints which the linguist believes to represent the competing factors supporting and penalizing the different candidate forms, similarity between the output distribution and the training data suggests that one's constraint set is sufficient. If, on the other hand, the constraint set cannot model the observed frequencies, this suggests that one is missing active constraints or that there are distinctions in the input that have gone unnoticed.

1.4 The Constraints of Aissen (1999) on Grammatical Relations

The stochastic OT analyses throughout will incorporate the constraints of Aissen (1999,2000) derived from the application of the harmonic alignment technique of Prince and Smolensky (1993) to syntax. In particular, the same constraints on the alignment of grammatical relations with thematic roles will be used in the accounts of both definiteness and person, and therefore will be explicated here. The technique of harmonic alignment is based on the principle that, given a binary structural prominence scale and a prominence scale on some other dimension X, elements which are prominent on dimension X will be attracted to structurally prominent positions, and elements which are non-prominent on dimension X will be attracted to structurally non-prominent positions (Aissen 1999). The formal definition is as follows:

(5) Alignment. Suppose given a binary dimension D_1 with a scale X > Y on its elements {X,Y}, and another dimension D_2 with a scale a > b …> z on its elements. The harmonic alignment of D_1 and D_2 is the pair of harmony scales:

H_x: X/a ≻ X/b ≻ … ≻ X/z
H_y: Y/z ≻ … ≻ Y/b ≻ Y/a

The constraint alignment is the pair of constraint hierarchies:

C_x: *X/z >> … >> *X/b >> *X/a
C_y: *Y/a >> *Y/b >> … >> *Y/z

(Prince and Smolensky 1993, p.136)

Aissen uses this technique to align the binary scale Su > Non-Su, a scale of structural prominence, with the thematic role hierarchy Agt > Pat.

(6) Grammatical Relations Prominence Scale: Su > Non-Su
(7) Thematic Role Prominence Scale Agent > Patient

The alignment of (6) with (7) produces the harmony scales in (8) and (9):

(8) Su/Agt ≻ Su/Pat

(9) Non-Su/Pat ≻ Non-Su/Agt

These harmony scales express the generalization that it is preferable to have agents rather than patients as subjects and preferable to have patients rather than agents as non-subjects. Inverting the harmony scales results in the constraint hierarchies in (10)-(11).

(10) *Su/Pat >> *Su/Agt
(11) *Non-Su/Agt >> *Non-Su/Pat

That is, it is a worse violation to have a patient subject than an agent subject, and a worse violation to have an agent non-subject than a patient non-subject. The constraints on non-subjects are converted into separate constraints penalizing objects and obliques as in (13)-(14), and the full set of constraints on the association of thematic roles with grammatical relations is in (12)-(14).

(12) *Su/Pat >> *Su/Agt
(13) *Obj/Agt >> *Obj/Pat
(14) *Obl/Agt >> * Obl/Pat

Since this constraint set will be used in the chapters to follow, its merits and drawbacks will be briefly discussed here. Firstly, one goal of Aissen's constraint set is to illustrate markedness of passives. Presumably, markedness of passives would be illustrated if the constraint set relating to semantic role penalized passive, and passive were driven by separate, discourse constraints. The subhierarchy in (12) seems to accomplish this goal. However, the constraint *Obj/Pat in (13) disfavors active without any higher-ranking constraint disfavoring passive. Thus, without any discourse constraints driving passive, one could have a language that disfavored actives and had only passives (by a high ranking of *Obj/Pat). Thus, it is unclear how Aissen's constraint set implies markedness of the passive. Secondly, as stated previously, in optimality theory it is assumed that the candidate set is infinite. Presumably, then, candidates include all possible assignments of semantic roles to grammatical relations. Consideration of this full candidate set produces potentially serious concerns for the constraint set in (12)-(14). For example, if one assumes that candidates realizing patients as obliques never win, since "true" patients in English are not generally realized as obliques, then the constraint *Obl/Pat must be ranked high, and *Obl/Agt must be ranked even higher due to the constraint subhierarchy *Obl/Agt >> *Obl/Pat, implying that the passive would never occur in English.

Here, these potential problems will simply be acknowledged. Following Aissen (1999), only the active and passive candidates will be considered. Constraints addressing semantic role will be used because there does appear to be a clear dispreference for passive in English, and without representing this constraint it is impossible to model the observed frequencies. However, for simplicity, the analyses presented will use only a subset of the constraints in (12)-(14). The constraints in (14) will be eliminated since they do not significantly differentiate the active and passive candidates beyond the constraints in (12)-(13).

In the next chapters I will discuss first the definiteness and then the person hierarchies. Each chapter will first discuss past proposed motivations for the hierarchy and relevant past work associating the hierarchy with voice. Then, the relevant constraints involving the hierarchy will be introduced and their predictions for frequencies of active and passive will be analyzed. Finally, a corpus study investigating the effects of the hierarchy on frequencies of passivization in English will be presented, its results analyzed, and the results of training the constraints on the frequency data in the stochastic OT framework reviewed.

Chapter 2

The Definiteness Hierarchy

That a hierarchy of definiteness has effects on syntax has been demonstrated by work on various phenomena in various languages, including differential object marking (Aissen 2000), split ergativity (Silverstein 1976), and subject and object selection (McFarland 1978, Givón 1979). In this chapter I present evidence of frequency effects of the definiteness hierarchy on the choice between active and passive in English. I begin by motivating why one would expect to see such effects by discussing various versions of and proposed groundings for the hierarchy and its links to subject and object selection. I go on to discuss Aissen's formalization of these ideas using the technique of harmonic alignment of prominence scales, and consider the predictions of her constraints in a stochastic optimality theory framework. Then I present the results of examining these predictions in a corpus study of actives and passives. It is demonstrated that the frequency effects observed are significant and in accordance with the predictions of the theory, and also that they support observations on the hierarchy made by Ariel (1990). Finally, the results of training the stochastic OT model on data obtained from a corpus are presented.

2.1 Explaining Definiteness Hierarchy Effects

Proposed hierarchies of definiteness have generally taken the approach of ranking noun phrase *forms*; thus they have ranked elements such as zero, pronoun, definite and indefinite. Below I review various ways in which the hierarchy of definiteness has been grounded and different forms it has taken, and how the various theories make predictions with regard to the interaction of the definiteness of agent and patient arguments with the likelihood of passivization.

Silverstein (1976) introduced a hierarchy of noun phrases based on features of both definiteness and animacy to account for split ergativity. In this hierarchy, pronouns were placed above proper nouns which in turn were placed above full noun phrases; the hierarchy was rooted in terms of the "inherent lexical content" of its elements. Along similar lines, Aissen (2000) roots the hierarchy "pronoun > proper noun > definite > indefinite specific > non-specific" in the degree to which the value assigned to the referent of the noun phrase is fixed. A shortcoming of these approaches is that the notion of "fixedness of lexical context" does not appear to be predictive of the phenomena in which the hierarchy will have effects. Silverstein also frames his hierarchy in terms of the likelihood of its elements to be agents; his claim is that elements higher on the hierarchy are most likely to be agents and those on the lower end most likely to be patients. The problems with this approach will be discussed further in the next chapter; however, here it is relevant that the notion of agentivity does not appear to accurately characterize the ranking of proper nouns higher than full nouns in Silverstein's hierarchy, since proper nouns, which can be inanimate, are ranked above full noun phrases, which can be animate.

Another class of approaches grounds the hierarchy in terms of information status. For example, the hierarchy "definite > indefinite" is analyzed by Givón (1976) as "merely a reflection of old information being the topic and new information being the assertion…" Chafe (1976) points out that definites can be both new and old; however, since indefinites are almost invariably new, this is still supportive of a hierarchy "definite > indefinite". Proper nouns are similar to definites in this respect (that is, they can be both old and new), while pronouns are invariably old. Thus, the distribution of old and new information supports the ordering "pronoun > proper noun/definite > indefinite".

A third class of approaches uses the notion of "accessibility". In Ariel's Accessibility Theory (1990), the use of particular referring expressions such as pronouns, proper nouns, definites, and indefinites is a strategy for marking the accessibility of the mental representation of discourse referents. (Ariel's theory is quite similar to that of Givón (1983), who uses the notion of scalar "topic accessibility" coded by a range of ranked grammatical devices.) Ariel argues that all referring expressions in all languages are arranged on a scale of accessibility, and the use of high accessibility referring expressions implies that the discourse referent has high accessibility to the addressee, while the use of low accessibility referring expressions implies that it has low accessibility to the addressee. Thus, the definiteness hierarchy can be interpreted as a ranking of "accessibility markers" from high accessibility markers (e.g. pronouns) to low accessibility markers (definites, proper nouns, indefinites).

Within all of these accounts there is a great deal of variation in the elements ranked and some disagreement in the relative ordering of certain elements. For example, some rank only "Definite > Indefinite" while others include separate categories for pronouns and proper nouns, and Ariel's scale includes agreement markers, zeros, distinctions between stressed and unstressed pronouns, and different types of proper nouns. In addition, approaches which rank proper nouns higher than definites conflict with the results of Ariel, who claims that the ordering between proper nouns and definites is not fixed, because of evidence that proper nouns do not class uniformly – last names or first names are more accessible than definite descriptions, which in turn are less accessible than full names. The close connection between proper nouns and definites is reflected in the fact that some languages use the definite article for proper nouns as well as common nouns (Chafe 1976); in English this tendency can be seen as well in proper nouns such as *The United States, the Sears Tower*.

The different groundings of the definiteness hierarchy have different predictions for the interaction of definiteness of agent and patient with the tendency to passivize. While certain elements having greater "inherent lexical content" does not seem to necessitate any interaction with passivization, rooting the hierarchy of definiteness in a tendency for higher elements to be discourse-old predicts that it will have an influence on passivization since placing old information in subject position and maintaining the old-before new pattern in discourse has been claimed to be one of the primary functions of the passive construction (cf. Birner and Ward 1998). Similarly, while Ariel (1990) does not connect the notion of accessibility to the choice of active and passive, if one assumed a tendency to place more high-accessibility elements in subject position and low-accessibility elements in non-subject positions, then the accessibility of agent and patient arguments would presumably affect the choice between active and passive. The association of elements higher on the hierarchy with subjects is supported by Keenan (1976), who states that "highly referential" NP's such as pronouns and proper nouns can always be subjects, by Givón (1979), who shows that subjects are usually definite, and by the fact that in a number of languages subjects cannot be non-specific (Manning 1996). The association of elements lower on the hierarchy with objects is supported by Keenan (1976), who cites Philippine languages in which objects cannot be definite (at least with non-relativized verbs (McFarland 1978)), and by the phenomenon of differential object marking, in which higher elements are always marked if lower elements are marked. Aissen (2000) characterizes this in terms of *markedness reversal* – the elements at the top of the hierarchy are unmarked as subjects but marked as objects, while the elements at the bottom are marked as subjects and unmarked as objects.

The aim here is to detect this markedness reversal in English by examining frequencies of active and passive. While it is clear in English that all combinations of definiteness in agent and patient are grammatical in both the active and the passive – that is, it is not ungrammatical to say *A girl was killed by him* or *A girl killed him* – presumably we might still observe, in the respective frequencies of active and passive for particular combinations of agent-patient accessibility, a tendency for certain combinations of definiteness in agent and patient to passivize less. In what follows, these ideas will be formalized using the constraints of Aissen (2000) and the Stochastic Optimality Theory framework of Boersma and Hayes (2001). Since the formalism of Aissen (2000) will be adopted, her version of the definiteness hierarchy will be used initially. However, the views of Ariel on the accessibility of proper nouns and definites will prove useful in the interpretation of the frequency results.

2.2 Formalizing the Effects of Definiteness On Voice

Aissen formalizes the markedness reversal between subject and object with the definiteness hierarchy using the technique of harmonic alignment of prominence scales described in the previous chapter. Aissen uses this technique to align the binary scale Su > Non-Su, a scale of structural prominence, with the definiteness hierarchy. As stated before, the definiteness hierarchy is represented as a prominence scale as in (2):

(2) Pronoun > Proper Noun > Definite > Indefinite Specific > Indefinite Non-Specific
(Pro > Proper > Def > IndefSpec > Non-Spec)

(3) Su > Non-Su

While Aissen uses the constraints resulting from harmonic alignment of these scales in conjunction with iconicity and economy constraints to account for differential object marking, there seems to be no reason why one might not use the constraints to model preferences for active and passive, (as in Aissen (1999) and discussed further in the next chapter).

Aligning the prominence scale in (2) with the scale in (3), in the same way as discussed in the previous chapter, we obtain the harmonic orderings shown in (4)-(5) and the constraint hierarchies shown in (6)-(7):

(4) Su/Pronoun ≻ Su/Proper≻ Su/Definite ≻ Su/Indefinite Specific ≻ Su/Non-Spec
(5) Non-Su/Non-Spec ≻ Non-Su/IndefSpec ≻ Non-Su/Definite ≻ Non-Su/Proper ≻ Non-Su/Pronoun

(6) *Su/Non-Spec >> *Su/IndefSpec >> *Su/Definite >> *Su/Proper >> *Su/Pronoun
(7) *Non-Su/Pronoun >> *Non-Su/Proper >> *Non-Su/Definite >> *Non-Su/IndefSpec >> *Non-Su/Non-Spec

By separating Non-Su into Object and Oblique, the final constraint set shown in (8)-(10) is obtained.

(8) *Su/Non-Spec >> *Su/IndefSpec >> *Su/Def >> *Su/Proper >> *Su/Pronoun
(9) *Obj/Pronoun >> *Obj/Proper >> *Obj/Def>> *Obj/IndefSpec >> *Obj/Non-Spec
(10) *Oblique/Pronoun >> *Oblique/Proper >> *Oblique/Def>> *Oblique/IndefSpec >> *Oblique/Non-Spec

Ideally, this constraint set should have the property of implying (all else being equal) that for any configuration in which the agent is of status X on the definiteness hierarchy and the patient is of status Y, if passive is obligatory, then passive will also be obligatory when the agent is of status X and the patient is of status $Z > Y$, or when the agent is of status $Z < X$ and the patient is of status Y. That is to say, if passive is obligatory at any square in the table below, then it will also be obligatory for all squares to the left and for all squares below.

Table I. All Possible Combinations of Definiteness in Agent and Patient

Agent ↓ Patient →	Pronoun	Proper Name	Definite	Indefinite-Spec	Non-Specific
Pronoun					
Proper name					
Definite					
Indefinite-Spec					
Non-Specific					

A short proof that the constraints introduced above do have this desired property is sketched below.

(11) Suppose passive is obligatory when agent is of definiteness status X and patient is of status Y. Then at least one of the three constraints penalizing the active is ranked higher than all the constraints penalizing the passive. So either (1) or (2) or (3). In all cases we have that any patient that is higher than Y on the definiteness hierarchy will also force passivization with an agent of status X, and similarly that any agent that is lower than X will force passivization with a patient of status Y:

1. *Obj/Y is ranked higher than *Su/Pat, *Su/Y, and *Oblique/X.
   Now if the agent remains at X but the patient is of status $Z > Y$, then *Obj/Z is ranked higher than *Obj/Y (by the constraint subhierarchy), so *Obj/Z is ranked higher than *Su/Pat and *Oblique/X (by transitivity). Also, *Su/Z is ranked lower than *Su/Y (by the constraint subhierarchy), so *Obj/Z is ranked higher than *Su/Z (by transitivity). So passive is obligatory.

Alternately, if the patient remains at Y but the agent is of status $Z < X$, then *Oblique/X is ranked higher than *Oblique/Z (by the constraint subhierarchy). So *Obj/Y is ranked higher than *Su/Pat, *Su/Y, and *Oblique/X (by transitivity).

2. *Su/X is ranked higher than *Su/Pat, *Su/Y, and *Oblique/X
   Now if the agent remains at X but the patient is of status $Z > Y$, then *Su/Z is ranked lower than *Su/Y (by c.s.), so *Su/X still outranks *Su/Pat, *Su/Z, and *Oblique/X (by transitivity). So passive is obligatory.
   Alternately, if the patient remains at Y but the agent is of status $Z < X$, then *Su/Z is ranked higher than *Su/X (by c.s.). So *Su/Z is ranked higher than *Su/Pat and *Su/Y (by transitivity). Also, *Oblique/Z must be lower than *Oblique/X (by c.s.). So *Su/Z is ranked higher than *Oblique/Z (by transitivity).

3. *Obj/Agt is ranked higher than *Su/Pat, *Su/Y, and *Oblique/X
   Now if the agent remains at X but the patient is of status $Z > Y$, then *Su/Z is ranked lower than *Su/Y (by c.s.). So *Agt/Obj continues to outrank *Su/Pat, *Su/Y, and *Oblique/Z (by transitivity).
   Alternately, if the patient remains at Y but the agent is of status $Z < X$, then *Oblique/Z is ranked lower than *Oblique/X (by c.s.). So *Obj/Agt continues to outrank *Su/Pat, *Su/Y, and *Oblique/Z (by transitivity).

So this constraint set has the desired property. It does not, however, seem to have the property of implying that we will see passivization specifically when the patient is of a higher definiteness than the agent. Rather, with this constraint set, it is entirely possible to have a language in which, given a definite agent, passivization is obligatory when the patient is an indefinite specific or higher, but active is obligatory when the patient is non-specific. This seems suited to modeling differential object marking, where marking of objects is independent of the definiteness status of the subject. However, it is unclear whether it is suited to passivization, which at least in the case of languages with categorical person-voice effects, has been analyzed as occurring specifically when the patient is lower on the person hierarchy than the agent (this will be discussed further in the next chapter). As will be seen, however, the properties of the constraint set are not problematic for modeling the frequency effects observed in English.

As mentioned before, in English, it is clear that all active and passive sentences, no matter what the configuration of definiteness of subject, object, and oblique, are grammatical. While *A boy was killed by her* may sound awkward, it is not ungrammatical, and does occur in certain discourse contexts (cf. Kato (1979), Utsugi (1998)). However, in a stochastic OT framework, the property of the constraint set shown above, (that if passivization is obligatory with a patient of status X and an agent of status Y, then it is also obligatory with a patient of status X and an agent of status $Z < Y$, or an agent of status Y and a patient of status $Z > Y$), translates into the following property: If passivization occurs at a certain frequency when the agent is of status X and the patient is of status Y, then it will occur at a higher or equal frequency when the agent is of status Z < X and the patient remains at status Y, and similarly it will occur at a higher or equal frequency when the agent remains at status X and the patient is of status Z > Y. This is briefly illustrated below by converting the proof in (11) into a Stochastic OT version:

(12) Suppose passive occurs at a frequency *f* when agent is of status X and patient is of status Y. Then at least one of the three constraints penalizing the active is ranked higher than all the constraints penalizing the passive *f % of the time*. So we have the following:

f % of the time one of *Obj/Y, *Su/X, and *Obj/Agt is ranked higher than *Su/Pat, *Su/Y, and *Obl/X.

Now if the agent remains at X but the patient is of status Z > Y, then *the normal distribution corresponding to *Obj/Z must have a mean greater than or equal to that of *Obj/Y* (by the constraint subhierarchy). Also, *the normal distribution corresponding to *Su/Z must have a mean lower than or equal to that of *Su/Y* (by the constraint subhierarchy). The other normal distributions remain the same. Therefore, *clearly passive must occur at a frequency greater than or equal to f %.*

Alternately, if the patient remains at Y but the agent is of status Z < X, then *the normal distribution corresponding to *Su/Z must have a mean greater than or equal to that of *Su/X*. Similarly, *the normal distribution corresponding to *Obl/Z must have a mean lower than or equal to that of *Obl/X*. The other normal distributions remain the same. Therefore, *clearly passive must occur at a frequency greater than or equal to f %.*

Therefore, despite the lack of grammaticality effects in English, we can still, given the set of inputs in the table above, expect to see progressively lower rates of passivization going left-to-right across each row, and progressively higher rates of passivization going top-to-bottom in each column. This testable hypothesis will be investigated below.

2.3 Evidence of Definiteness-Voice Interactions

Categorical definiteness-voice interactions have been observed in Lummi, Lushootseed, Squamish and Chamorro. In all of these languages, active sentences are excluded when the agent is nominal and the patient is pronominal (Jelinek and Demers 1983, Cooreman 1987). In this situation, the passive must be used. This could be accounted for by ranking *Obj/Pronoun above all the constraints penalizing passives with nominal agents and pronominal patients (which in this case would be *Su/Pat, *Obl/Proper,*Obl/Definite, *Obl/Indefinite, and *Su/Pronoun). This ranking is shown in the table below.

Table II.

/Nominal Agent – Pronominal Patient/	*OBJ/PRONOUN	*OBLIQUE/(DEF/INDEF/PROPER)	*SU/PAT	*SU/PRONOUN
☞ [Passive]		*	*	*
*[Active]	*!			

Alternately, ranking constraints penalizing actives with nominal subjects (that is *Su/Def, *Su/Indef, and *Su/Pronoun) highest would produce the same effect.

Frequentistic definiteness-voice effects in English have been demonstrated by Givón (1979), who shows that indefinite subjects in English main clause active declarative sentences occur at a quite low frequency – approximately 10% of English subjects are indefinite, as opposed to 90% definite. Francis et al (1999) show in a study of the Switchboard corpus (English conversation) that among subjects, 91% are pronominal while only 9% are lexical, while among objects 66% are lexical and 34% pronominal. Estival and Myhill (1988) demonstrate that pronominal agents are less likely to passivize (0%) than nominal agents (5%), and that definite agents are less likely to passivize (1%) than indefinite agents (4%). They also show that pronominal patients are more likely to passivize (17%) than nominal patients (5%), and definite patients more likely to passivize (12%) than indefinite patients (4%). Svartvik (1966) finds consistently across three texts (M_1, M_2, and J_1) that the proportion of pronouns in subject position of passives is much higher than the proportion of pronouns in object position of actives (66% vs. 25% in M_1, 66% vs. 22% in M_2, and 25% vs. 2% in J_1). Similarly, the proportion of pronouns in subject position of actives is much higher than the proportion of pronouns in *by*-phrases of passives (22%, 66%, and 33% versus approximately 2%). These results demonstrate a frequency effect of markedness reversal. Ransom (1979), using the hierarchy "definite-referential > indefinite-referential > indefinite nonreferential" finds that 44% of English passives have subjects higher on the definiteness hierarchy than agents, 47% have subjects equally high, and only 9% have subjects lower. Thus, there is preliminary evidence that definiteness influences the choice between active and passive in English. Below, I describe a study aimed specifically at testing the frequency gradation predictions of the constraints resulting from harmonic alignment of the definiteness and relational scales in a stochastic OT model.

2.4 Effects of the Definiteness Hierarchy on Frequencies of Passivization in English

2.4.1 Methodology

The corpus used was the Wall Street Journal Corpus, a sub-corpus of the Penn Treebank (Marcus et al. 1993). The WSJ corpus consists of a million words of 1989 Wall Street Journal newswire, fully parsed and annotated. This corpus was chosen since it was assumed it would have a larger number of third person pronouns and proper nouns than a corpus of conversation. Also, the higher rates of passivization in the WSJ corpus would allow for a better comparison of differences in the tendencies to passivize of different agent-patient definiteness combinations. The WSJ corpus can be searched easily using the tgrep program, which allows the user to specify a pattern for the tree structure of a sentence, and then returns all the trees in the corpus corresponding to that pattern. The goal was to find the numbers of active and passive outputs in the corpus corresponding to all the combinations of definiteness in agent-patient pairs. Due to the difficulty of automatically differentiating the indefinite specific and nonspecific categories, these were collapsed into the single category indefinite. This left the following sixteen inputs:

1. /pronoun agent + pronoun patient/
2. /pronoun agent + proper noun patient/
3. /pronoun agent + definite patient/
4. /pronoun agent + indefinite noun patient/

5. /proper noun agent + pronoun patient/
6. /proper noun agent + proper noun patient/
7. /proper noun agent + definite patient/
8. /proper noun agent + proper noun patient/

9. /definite agent + pronoun patient/
10. /definite agent + proper noun patient/
11. /definite agent + definite patient/
12. /definite agent + indefinite patient/

13. /indefinite agent + pronoun patient/
14. /indefinite agent + proper noun patient/
15. /indefinite agent + definite patient/
16. /indefinite agent + indefinite patient/

Due to the difficulty of defining the notions of agent and patient and the even greater difficulty of automatically detecting them in the corpus, these notions were approximated as the logical subjects and objects of transitive verbs. That is, any transitive verb was assumed to have agent and patient arguments, with its agent corresponding to its subject in an active sentence and its patient corresponding to its object. In this sense our notion of agent and patient is better characterized as proto-agent and proto-patient (assuming that Dowty (1991) is correct in theorizing that any transitive verb will have proto-agent mapped to subject and proto-patient mapped to object). Thus, for the first input the script would detect the number of active sentences with pronoun subjects and pronoun objects and the number of passive sentences with pronoun subjects and pronouns in the oblique. Only full *by*-phrase passives were counted, since it would be difficult to determine which of inputs (1)-(16) an agentless passive corresponded to (due to the absence of the agent argument).

The main methodological issue was how to detect each kind of noun phrase; that is pronoun, proper noun, definite, and indefinite. In the corpus, nouns are annotated “NN” for singular noun and “NNS” for plural noun;

proper nouns are annotated "NNP" for singular proper nouns and "NNPS" for plural proper nouns, pronouns are annotated "PRP" (possessive pronouns such as *her* are annotated differently as PRP$), and determiners are annotated "DT". However, at the level of noun phrases it can still be complicated differentiating the various kinds of noun phrases. Definites, for example, cannot simply be detected as those noun phrases containing the determiner *the* since this would include proper nouns such as *The United States*. Additionally, attempting to automatically differentiate different types of noun phrases is complicated by the fact that the distinctions between types can be fuzzy. In the case of *the Kent cigarettes*, for example, it is unclear whether to call this a definite or a proper noun. In the case of *the Honda* or *the Wurlitzer*, presumably one would want to call these definites even though they have exactly the same form as t*he United States*.

For simplicity, simple definitions of definite and indefinite were used. Definites were detected as those noun phrases whose leftmost daughter was one of the determiners *the*, *this*, *that*, *these,* or *those* and did not have a sister which was a proper noun (to exclude *the United States*.) Similarly, indefinites were detected as those noun phrases whose leftmost daughter was one of the determiners *a*, *an*, or *some* and did not have a sister which was a proper noun. Pronouns were detected as those noun phrases whose leftmost daughter was a word dominated by "PRP". Proper noun phrases were detected as those noun phrases whose leftmost daughter was a proper noun or the determiner *the* followed by a proper noun and which did not have as a sister a common noun (to exclude *the Wyoming area*). All "possessed" noun phrases were excluded (e.g. *Mary's hat*) for simplicity. The scripts for detecting each type of noun phrase are reproduced on p. 32 of the appendix.

We were interested in isolating the effect of the person of the agent and patient arguments on the realization of the inputs. That is, we sought to answer the question: *all else being equal*, does definiteness have an effect on the probability of passivization? For this reason we tried to exclude inputs in which the choice between active and passive was dictated or influenced by other factors. This rationale dictated several methodological decisions. Firstly, only main verbs were considered. For example, *I killed Mary* would be counted as an active sentence with pronoun subject and proper noun object, but *John told me to kill Mary* or *John told Mary that I killed Susan* would not. This was to avoid cases in which the main clause verb would dictate the choice between active and passive. For example, it is impossible to express *John told me to kill Mary* with a passive in the subordinate clause. Secondly, sentences whose main verbs were judged not to have a corresponding passive form (as with *have*) were not counted. The justification for this was that in these cases the passive candidate would be eliminated due to the absence of a passive form. This additional factor would complicate the results. For example, if first persons had a greater tendency to appear as subjects of the verb *have* (which does not passivize), then this would artificially create the impression that first persons did not passivize, when in fact the mediating factor would be not the person of the subject argument but the verb itself. Therefore, a list of nonpassivizing verbs was compiled by first producing a list of all the verbs appearing in active and passive sentences and then removing from that list all the verbs which were judged not to passivize. Since a number of verbs have multiple meanings, some of which passivize while others do not (e.g. *weigh* can passivize in *He was weighed by the doctor* but not in *Ninety-eight pounds were weighed by him*) such cases were determined by taking a sample of sentences in which this verb appeared and judging whether the majority of the senses in that sample passivized or not. The list of verbs judged not to passivize (in the majority of their senses) appears on p.36 and the list of verbs judged to passivize appears on pp.37-44. Additionally, sentences with main verb *born* (e.g. *Mary was born in 1968*) were not counted, since no agent is possible with *born* (thus one could not have *Mary was born by her mother in 1968*.) Thirdly, empty subjects (as in imperatives) were thrown out since the constraints of Aissen address overtly expressed arguments. Sentences containing expletive subjects were thrown out since expletive subjects arguably do not correspond to any semantic role and thus cannot be classed as corresponding to any of the inputs in (1)-(4) (also they do not passivize). Finally, all sentences containing coordinated subjects, coordinated logical subjects, and coordinated objects were removed to avoid cases in which one conjunct had one definiteness status and the other had another. Such coordination of arguments differing in definiteness would make it unclear which of inputs (1)-(16) a sentence corresponded to.

We used the tgrep program for searching the Penn Treebank by specifying desired tree patterns. The patterns were specified as follows. Subjects are marked in the corpus as “NP-SBJ”, logical subjects of passive sentences are marked “NP-LGS”, and objects were approximated as the first noun phrase sister of the verb. (The marking of “NP-LGS” is quite consistent in the corpus; a quick search shows that it is missing in only one case.) Thus *by*-phrase passives were counted as those sentences containing an "NP-LGS" in a prepositional phrase sister of a verb of form “VBN” (past participle). Active sentences were those sentences whose main verb had a non-empty object (empty objects could not be counted since passive sentences are annotated as having empty objects, or traces). Agentless passives were counted as those sentences with a past participle verb (annotated “VBN”), which did not have a PP sister containing a logical subject, but were dominated by a VP which had a sister of the form *be* or *get*.

Topicalized sentences were also counted, e.g. *Bears, I like*, or *Booth, Lincoln was killed by*. Topicalized elements are annotated “NP-TPC” and topicalized active sentences were detected as those sentences whose “NP-SBJ” had a sister “NP-TPC” and whose main verb was sister to an empty noun phrase. This method of detecting topicalizations is liable to also detect passive sentences or sentences such as *Mary, I gave a book*, so that the results must be hand-filtered to include only actives where the object is topicalized, or passives where the passive agent is topicalized. Since the number of topicalizations is quite small (under 5), this is not difficult.

Topicalized *by*-phrase passives were detected as those sentences whose “NP-SBJ” had a sister “NP-TPC” and whose “NP-LGS” dominated an empty noun phrase (marked “-NONE-”). Presumably, topicalized agentless passives could not occur. The total number of actives corresponding to any particular input was counted as the sum of the topicalized and non-topicalized actives for that input, and similarly for the passives.

Due to the increasing length of the tgrep commands (in particular, eliminating non-passivizable verbs involved specifying a list of over 100 verbs), it became impractical to hand-enter them. Therefore, the commands were entered into a PERL script which, when run, issued the commands and printed the results (the number of actives and passives found for each input). The first PERL script (on p.33) is one of the scripts detecting actives corresponding to each of the four inputs. Notice that this script detects verbs whose second daughter is a noun phrase. This is because it appears to be impossible to tell tgrep to find the first noun phrase daughter of the VP. Thus, to find the first noun phrase daughter, one must run separate scripts in which the first noun phrase daughter corresponds to the second daughter, the third daughter, and so on. This was done for the second through eighth daughters (there are not any cases in which the first NP is the eighth daughter, and we assume the same would be true for all daughters greater than eight as well). By adding up the trees produced by each of these scripts, we obtain a total number of non-topicalized actives. The second PERL script (on p.34) found non-topicalized *by*-phrase passives corresponding to each of the four inputs. The third PERL script (on p.35) found topicalized actives and the fourth PERL script (on p.36) found topicalized passives corresponding to each of the four inputs.

2.4.2 Results

Table I. Raw Data

Agent ↓ Patient →	Pronoun	Proper Noun	Definite	Indefinite
Pronoun	A: 103	A: 80	A: 264	A: 262
	P: 0	P: 0	P: 0	P: 0
Proper Noun	A:68	A: 190	A:486	A: 736
	P: 8	P: 21	P: 48	P: 10
Definite	A:52	A:77	A:387	A:450
	P: 12	P: 7	P: 30	P: 5
Indefinite	A:19	A:20	A:70	A:86
	P:8	P:11	P:28	P:7

Table II. Rates of Passivization

Agent ↓ Patient →	Pronoun	Proper Noun	Definite	Indefinite
Pronoun	0	0	0	0
Proper Noun	10.5	10.0	9.0	1.3
Definite	18.8	8.3	7.2	1.1
Indefinite	29.6	35.5	28.6	7.5

The raw data is presented in Table I, and the rates of passivization presented in Table II are calculated from Table I as the percentage 100*passives/(actives+passives). On the next page I present the results of the Fisher Exact Test for determining whether differences between two boxes are significant. The Fisher Exact Test is considered more accurate, though more difficult to calculate, than the more familiar t-test. I use $p < 0.05$ as the test for significance. Significance is calculated for every pair of boxes which fall in the same row or the same column, and a significant difference between two boxes is indicated by a line drawn between them. The hypothesis was that rates of passivization would either decrease or remain the same from left-to-right across rows and increase or remain the same from top-to-bottom in columns.

Table III. Significance Between Boxes in the Same Row

A ↓ P →	Pronoun	Proper Noun	Def	Indef
Pronoun	A: 103 P: 0	A: 80 P: 0	A: 264 P: 0	A: 262 P: 0
Proper Noun	A:68 P: 8	A: 190 P: 21	A:486 P: 48	A: 736 P: 10
Definite	A:52 P: 12	A:77 P: 7	A:387 P: 30	A:450 P: 5
Indef	A:19 P:8	A:20 P:11	A:70 P:28	A:86 P:7

Table IV. Significance Between Boxes in the Same Column

A ↓ P →	Pronoun	Proper Noun	Definite	Indefinite
Pronoun	A: 103 P: 0	A: 80 P: 0	A: 264 P: 0	A: 262 P: 0
Proper Noun	A:68 P: 8	A: 190 P: 21	A:486 P: 48	A: 736 P: 10
Definite	A:52 P: 12	A:77 P: 7	A:387 P: 30	A:450 P: 5
Indefinite	A:19 P:8	A:20 P:11	A:70 P:28	A:86 P:7

In the first row, there were no statistically significant differences (since they were all zeros). In the second and fourth rows, there were statistically significant differences between the first three boxes and the fourth, but no others. In the third row, there were statistically significant differences between every pair of boxes except the definite agent-proper noun patient box and definite-agent-definite patient box. In the first column (pronoun patient), there were statistically significant differences between the pronoun agent category and all other agent categories, and between the proper noun agent category and the indefinite agent category, but not between the definite and indefinite agents or the proper noun and definite agents. In the second column (proper noun patient), the only difference which was not significant was between the definite agent and proper noun agent categories. This was also the case in the third column and the fourth columns, except that in the fourth column the difference between the pronoun and definite agents was not significant.

The results show a clear interaction between the choice of passive and the definiteness of agent and patient arguments. Every statistically significant difference (31 out of 48 pairs of boxes) was in the direction predicted by the theory. There were no statistically significant differences in a direction opposite to the predictions of the theory. There was also no statistically significant difference in the behavior of proper nouns and definites. Thus, as predicted by the theory, the frequency of passivization either decreases or remains the same as we go from left to right in each row, and increases or stays the same as we go from top to bottom in each column. The lack of significance in comparing adjacent proper noun-definite boxes supports Ariel's noncategorical ranking of proper nouns and definites. In fact, when collapsing proper nouns and definites into a single category, all column differences are significant using the Fisher Exact Test, and excepting the first row (where all boxes are zero) and the comparison between indefinite agent – pronoun patient and indefinite agent – proper noun/def patient, all row differences are significant.

The frequency data also attests to a large "distance" between indefinites and the other elements on the hierarchy. In particular, the data in the rows and in the second through fourth columns suggests that the jump between different elements on the hierarchy is not equal; in these cases the rate of passivization of definites and proper nouns is far closer to that of pronouns than that of indefinites.

2.5 Stochastic OT Analysis

We now consider the results of attempting to model the observed frequencies using the constraints presented previously and the gradual learning algorithm. Since the categories of indefinite specific and non-specific were collapsed into one, there remain twelve constraints in three subhierarchies from the alignment of the definiteness hierarchy with the grammatical relations hierarchy. These are reproduced below:

(13) *Su/Indef >> *Su/Def >> *Su/Proper >> *Su/Pronoun
(14) *Obj/Pronoun >> *Obj/Proper >> *Obj/Def>> *Obj/Indef
(15) *Oblique/Pronoun >> *Oblique/Proper >> *Oblique/Def>> *Oblique/Indef

In addition, we have the constraints resulting from alignment of the semantic role hierarchy Agt > Pat with the grammatical relations hierarchy:

(16) *Su/Pat >> *Su/Agt
(17) *Obj/Agt >> *Obj/Pat

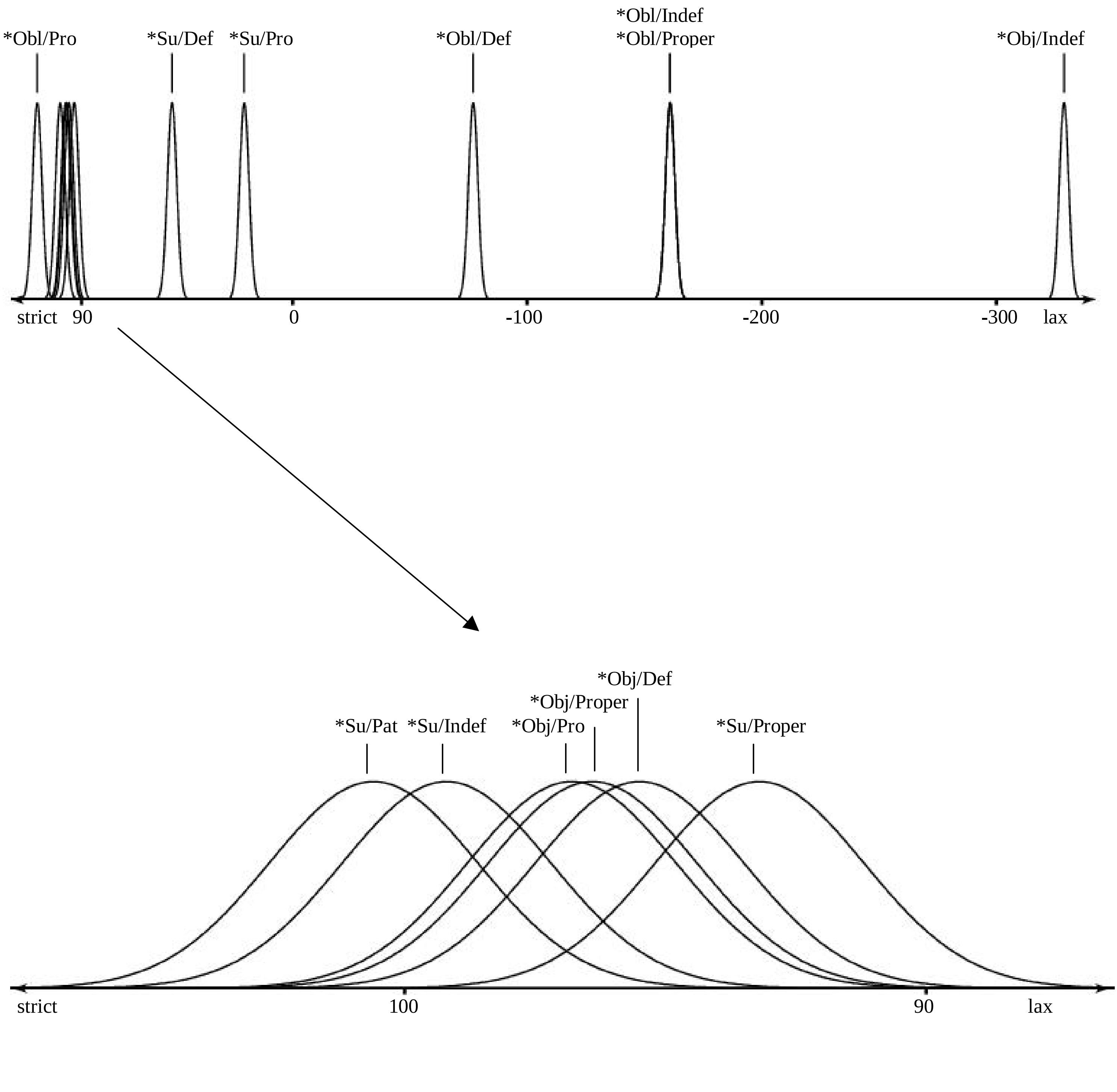
(The constraints on obliques are ignored here for simplicity since they do not further differentiate the candidates.)

The gradual learning algorithm was then used to train the above sixteen constraints on the data obtained from the Wall Street Journal Corpus. In training, one can set the algorithm to respect subhierarchies. The algorithm was set to respect the subhierarchy "Pronoun > Definite/Proper Noun > Indefinite". That is, no hierarchy between proper nouns and definites was encoded. After training, the following constraint rankings were in place:

Table V. Constraint Rankings After Training

Constraint	**Rank After Training**
*Su/Indef	99.2
*Su/Def	51.5
*Su/Proper	93.2
*Su/Pro	20.7
*Obj/Pro	96.8
*Obj/Proper	96.4
*Obj/Def	95.5
*Obj/Indef	-328.8
*Oblique/Pro	109.0
*Oblique/Proper	-160.9
*Oblique/Def	-76.9
*Oblique/Indef	-160.9
*Su/Pat	100.6
*Su/Agt	92.4
*Obj/Agt	100.0
*Obj/Pat	95.5

These rankings are depicted graphically below (to avoid overcrowding only a few are labeled and the semantic role constraints are left out, excepting *Su/Pat):



These constraint rankings produce the following percentages of passive. The figures in parentheses represent the original distribution:

Table VI. Output Distributions of Constraint Rankings in Table V.

Agent ↓ Patient →	Pronoun	Proper Noun	Definite	Indefinite
Pronoun	**0** (0)	**0** (0)	**0** (0)	**0** (0)
Proper Noun	**11.6** (10.5)	**9.6** (10.0)	**7.0** (9.0)	**1.6** (1.3)
Definite	**11.3** (18.8)	**9.5** (8.3)	**6.9** (7.2)	**1.6** (1.1)
Indefinite	**34.6** (29.6)	**34.1** (35.5)	**32.8** (28.6)	**3.8** (7.5)

The output distribution matches the input distributions to a reasonable degree. The constraint ranking following training does not produce any differentiation between proper noun and definite *agent* arguments (thus the second and third rows look essentially the same). The algorithm does however produce a small difference between the proper noun and definite patient arguments in the second, third and fourth rows. This may be due to the consistently higher rate of passive in the proper noun column over the definite column in the original data. Even though these differences were shown not to be significant, it is important to note that when one enters a pair distribution for training with the gradual learning algorithm, this pair distribution is trained on several times. Thus, the data the algorithm sees in learning is liable to contain significant differences between outputs for different inputs that were not significant in the data obtained from the corpus, and can drive the algorithm to reproduce these differences. Presumably the differences in the proper noun and definite rows are not reproduced because they contradict one another **-** as can be seen below, in the original data the definite rate exceeds the proper noun rate in the first column but the reverse holds for the last three columns.

Proper Noun	**11.6** (10.5)	**9.6** (10.0)	**7.0** (9.0)	**1.6** (1.3)
Definite	**11.3** (18.8)	**9.5** (8.3)	**6.9** (7.2)	**1.6** (1.1)

It should also be noted that the algorithm cannot reproduce those differences which contradict the subhierarchies. Thus, the higher rate of passivization for /indefinite-agent+proper-noun-patient/ compared to /indefinite-agent+pronoun-patient/ is not reproduced by the algorithm, but instead the two are quite close together in the output distribution.

Consider how the constraint rankings resulting from training produce the output distribution in Table VI. The high ranking of *Oblique/Pronoun accounts for the zero rate of passivization in the first row. At 109, *Oblique/Pronoun is the highest-ranked constraint and over ten units higher than any constraint favoring the passive when the agent is a pronoun. Therefore, the odds of it being outranked by a constraint favoring the passive are approximately 1 in 5000 or less (cf. section 1.3.2). It is unclear whether less than 1 in 5000 is equivalent to "ungrammatical". If it were, our constraint rankings would clearly be flawed because passives with pronoun agents are grammatical in English (cf. section 2.2). However, certain improvements are necessary in our model in any case which could preempt this question. Firstly, topicality constraints (which have been hypothesized to play the primary role in driving the English passive) are missing from our constraint set. A ranking of a topicality constraint close to the *Oblique/Pronoun constraint could drive passive with a pronoun agent when the patient was highly topical. Secondly, our data has no instances of a pronoun agent in a *by*-phrase. Presumably a language learner exposed to millions of sentences *would* encounter such instances and this would drive down the *Oblique/Pronoun constraint.

Now consider the parallel rates of passivization in the first three columns of the second and third rows (the shaded area in the table below).

Agent ↓ Patient →	Pronoun	Proper Noun	Definite	Indefinite
Pronoun	**0**	**0**	**0**	**0**
Proper Noun	**11.6**	**9.6**	**7.0**	**1.6**
Definite	**11.3**	**9.5**	**6.9**	**1.6**
Indefinite	**34.6**	**34.1**	**32.8**	**3.8**

The constraints relevant to these six inputs are *Obj/Pronoun, *Obj/Proper, *Obj/Def, *Su/Proper, all the grammatical role constraints, and *Su/Def, *Su/Pronoun, *Obl/Proper , *Obl/Def. Due to the low ranking of the latter four constraints (ranked at 51.5, 20.7, -160.9,-76.9), they play no role in determining the frequency of passivization. With the grammatical role constraints alone, one would see the same rate of passive for all of these inputs – approximately 4% (it is not 0% because of *Obj/Pat and *Su/Agt, which are ranked in the nineties and penalize the active). However, with the high ranking of *Object/Pronoun, *Object/Def, and *Object/Proper, (ignoring *Obj/Agt, which nothing violates, these are the fourth through sixth highest constraints at 96.8, 96.4, and 95.5, respectively) we have higher rates of passivization of 7%-12%. These decrease progressively moving left to right in the shaded area due to the progressively lower ranking of the *Obj/Pronoun, *Obj/Def, and *Obj/Proper constraints. The slightly higher rate of passivization in the second row compared to the third row may be due to

*Su/Proper, which ranked at 93.2 can still play a role, penalizing active with proper noun agents and penalizing passive with proper noun patients.

Now consider the last row, starting with the three leftmost squares. The overall high rate of passivization is due to the high ranking of *Su/Indef (penalizing active) at 99.2, the third-highest constraint. However, *Su/Pat is ranked higher at 100.6, meaning that the rate of passive is suppressed. With these two constraints alone, the rate of passive would be 31% ; however, the object constraints still play a role, driving up the rate of passive. In the bottom right square, *Su/Indef plays no role because both active and passive violate it. Thus, the rate of passive is the 4% which results from the grammatical role constraints alone. Finally, the 1.6% rates of passive in the indefinite patient column is due to the combination of the high-ranked *Su/Indef and *Su/Pat constraints, both of which penalize passive in this case.

2.6 Conclusion

In this chapter it has been demonstrated that the definiteness hierarchy has effects on the frequency of passivization in English, and that these effects are consistent with the frequency-gradation predictions of constraints resulting from harmonic alignment of the person and grammatical relations hierarchies (as in Aissen 1999) in a Stochastic Optimality Theory framework. It has also been shown that these frequencies can be modeled within Stochastic Optimality Theory. In the next chapter, I examine similar predictions with regard to the hierarchy of person.

Appendix

```
#Script Definitions for the Different Types of Noun Phrases:

$defArt="the|The|this|This|that|That|these|These|those|Those";
$indefArt="a|an|A|An|some|Some";
$anyNode="/DT/|/NN/|/NP/|/PRP/|/WP/|JJ|CD";

#Scripts for each type of noun phrase. These scripts are then interpolated into the scripts for active and passive.

#Definite noun phrases not involving possession:
$defNP = " <<, ( $defArt > (DT !\$ /NNP/) !>> ADJP|QP|NAC !>> (/^NP/ \$ CC \$ /^NP/) !>> ($anyNode \$ POS) !>> (NP \$ NP \$ \\,))";

#Indefinites
$indefNP = " <<, ( $indefArt > (DT !\$ /NNP/) !>> ADJP|QP|NAC !>> (/^NP/ \$ CC \$ /^NP/) !>> ($anyNode \$ POS) !>> (NP \$ NP \$ \\,))";

#Proper Nouns
$properNP = " <<, (__ > (DT|/NNP/ > (/NP/ !< NN|NNS !< (DT !< the|The) < /NNP/)) !>> ADJP|QP|NAC !>> (/^NP/ \$ CC \$ /^NP/) !>> ($anyNode \$ POS) !>> (NP \$ NP \$
\\,))" ;

#Pronouns
$pronounNP = " <<, (__ !>> (/^NP/ \$ CC \$ /^NP/) > PRP !>> (NP \$ NP \$ \\,))";
```

```
#________________________________________________________________
#        Commands for counting active sentences
#________________________________________________________________


#Pronoun Subjects
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $pronounNP )) >> (S < (/NP-SBJ/ $pronounNP))!>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>(/S/>>/S/)\">out1");
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $properNP )) >> (S < (/NP-SBJ/ $pronounNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \">out2");
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $defNP )) >> (S < (/NP-SBJ/ $pronounNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >out3");
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $indefNP )) >> (S < (/NP-SBJ/ $pronounNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \">out4");

#Proper Noun Subjects
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $pronounNP )) >> (S < (/NP-SBJ/ $properNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \">out5");
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $properNP )) >> (S < (/NP-SBJ/ $properNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >out6");
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $defNP )) >> (S < (/NP-SBJ/ $properNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >out7");
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $indefNP )) >> (S < (/NP-SBJ/ $properNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >out8");


#Definite Subjects (and every kind of object)
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $pronounNP )) >> (S < (/NP-SBJ/ $defNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >out9");
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $properNP )) >> (S < (/NP-SBJ/ $defNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >out10");
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $defNP )) >> (S < (/NP-SBJ/ $defNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >out11");
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $indefNP )) >> (S < (/NP-SBJ/ $defNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >out12");

#Indefinite Subjects
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $pronounNP )) >> (S < (/NP-SBJ/ $indefNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \">out14");
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $properNP )) >> (S < (/NP-SBJ/ $indefNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >out13");
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $defNP )) >> (S < (/NP-SBJ/ $indefNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >out15");
system ("tgrep -wi \"/VB/ !< ($string) > (VP <2 (NP $indefNP )) >> (S < (/NP-SBJ/ $indefNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >out16");
```

```
#__________________________________________________________
#        Commands for counting passive sentences
#__________________________________________________________

#Pronoun Agents (and every type of patient)
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $pronounNP ))) >> (S < (/NP-SBJ/ $pronounNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \"
>pass1 ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $pronounNP ))) >> (S < (/NP-SBJ/ $properNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \"
>pass2 ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $pronounNP ))) >> (S < (/NP-SBJ/ $defNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \"
>pass3 ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $pronounNP ))) >> (S < (/NP-SBJ/ $indefNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \"
>pass4 ");

#Proper Noun Agents (and every type of patient)
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $properNP ))) >> (S < (/NP-SBJ/ $pronounNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \"
>pass5 ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $properNP ))) >> (S < (/NP-SBJ/ $properNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \"
>pass6 ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $properNP ))) >> (S < (/NP-SBJ/ $defNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >pass7
");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $properNP ))) >> (S < (/NP-SBJ/ $indefNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \"
>pass8 ");


#Definite Agents (and every type of patient)
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $defNP ))) >> (S < (/NP-SBJ/ $pronounNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \"
>pass9 ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $defNP ))) >> (S < (/NP-SBJ/ $properNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \"
>pass10 ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $defNP ))) >> (S < (/NP-SBJ/ $defNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >pass11
");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $defNP ))) >> (S < (/NP-SBJ/ $indefNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >pass12
");

#Indefinite Agents
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $indefNP ))) >> (S < (/NP-SBJ/ $pronounNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \"
>pass13 ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $indefNP ))) >> (S < (/NP-SBJ/ $properNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \"
>pass14 ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $indefNP ))) >> (S < (/NP-SBJ/ $defNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >pass15
");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ $indefNP ))) >> (S < (/NP-SBJ/ $indefNP )) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \"
>pass16 ");
```

```
#___________________________________________________________
#	Commands for counting topicalized active sentences
#___________________________________________________________

#Pronoun Agents and Every Type of Patient
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $pronounNP) < (/NP-SBJ/ $pronounNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top1.txt");
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $properNP) < (/NP-SBJ/ $pronounNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top2.txt");
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $defNP) < (/NP-SBJ/ $pronounNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top3.txt");
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $indefNP) < (/NP-SBJ/ $pronounNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top4.txt");

#Proper Noun Agents and Every Type of Patient
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $pronounNP) < (/NP-SBJ/ $properNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top5.txt");
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $properNP) < (/NP-SBJ/ $properNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top6.txt");
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $defNP) < (/NP-SBJ/ $properNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top7.txt");
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $indefNP) < (/NP-SBJ/ $properNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top8.txt");

#Definite Agents and Every Type of Patient
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $pronounNP) < (/NP-SBJ/ $defNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top9.txt");
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $properNP) < (/NP-SBJ/ $defNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top10.txt");
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $defNP) < (/NP-SBJ/ $defNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/)\"
>top11.txt");
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $indefNP) < (/NP-SBJ/ $defNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top12.txt");

#Indefinite Agents and Every Type of Patient
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $pronounNP) < (/NP-SBJ/ $indefNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top13.txt");
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $properNP) < (/NP-SBJ/ $indefNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top14.txt");
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $defNP) < (/NP-SBJ/ $indefNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top15.txt");
system ("tgrep -wi \"/VB/ !< ($string)  > (VP < (NP </NONE/)) >> (S< (/NP-TPC/ $indefNP) < (/NP-SBJ/ $indefNP )  ) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>>
(/S/>>/S/)\" >top16.txt");
```

```
#____________________________________________________________
#	Commands for counting topicalized passive sentences
#____________________________________________________________

#Pronoun Agents and Every Type of Patient
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $pronounNP) < (/NP-SBJ/ $pronounNP)) !>>
/PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >top17.txt ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $pronounNP) < (/NP-SBJ/ $properNP)) !>>
/PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >top18.txt ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $pronounNP) < (/NP-SBJ/ $defNP)) !>>
/PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >top19.txt ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $pronounNP) < (/NP-SBJ/ $indefNP)) !>>
/PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >top20.txt ");

#Proper Noun Agents and Every Type of Patient
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $properNP) < (/NP-SBJ/ $pronounNP)) !>>
/PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >top21.txt ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $properNP) < (/NP-SBJ/ $properNP)) !>>
/PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >top22.txt ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $properNP) < (/NP-SBJ/ $defNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/
!>> (/S/>>/S/) \" >top23.txt ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $properNP) < (/NP-SBJ/ $indefNP)) !>>
/PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >top24.txt ");

#Definite Agents and Every Type of Patient
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $defNP) < (/NP-SBJ/ $pronounNP)) !>>
/PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >top25.txt ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $defNP) < (/NP-SBJ/ $properNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/
!>> (/S/>>/S/) \" >top26.txt ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $defNP) < (/NP-SBJ/ $defNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/
!>> (/S/>>/S/) \" >top27.txt ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $defNP) < (/NP-SBJ/ $indefNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/
!>> (/S/>>/S/) \" >top28.txt ");

#Indefinite Agents and Every Type of Patient
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $indefNP) < (/NP-SBJ/ $pronounNP)) !>>
/PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >top29.txt ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $indefNP) < (/NP-SBJ/ $properNP)) !>>
/PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/ !>> (/S/>>/S/) \" >top30.txt ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $indefNP) < (/NP-SBJ/ $defNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/
!>> (/S/>>/S/) \" >top31.txt ");
system("tgrep -wi \"(VBN !< $string) > (VP < (PP < (/NP-LGS/ < /NONE/))) >> (S < (/NP-TPC/ $indefNP) < (/NP-SBJ/ $indefNP)) !>> /PP/|/NP/|/FRAG/|/X/|/UNF/|/AD/|/EDIT/
!>> (/S/>>/S/) \" >top32.txt "
```

Verbs Judged to be Nonpassivizing in the Majority of Their Senses

agreed
averaged
averaging
bear
bears
born
borne
coextrude
constitute
contain
contained
contains
cost
costing
costs
culminates
emerged
go
goes
gone
had
Had
has
have
having
hemorrhaging
involve
involved
involves
is
lagged
lasted
lasts
looked
mince
nearing
quit
remains
rise
rose
rumored
soared
stand
surged
talked
talking
total
totals
tumbled
waited
walk
want
wanted
wants
went

Verbs Included

abandoned
abandoning
abandons
abdicate
abolished
absorb
absorbed
absorbs
accelerate
accelerated
accelerating
accept
accepted
accepts
accompanied
accompanies
accomplished
accrue
accused
accusing
accustomed
achieve
achieved
acknowledge
acknowledged
acknowledges
acquire
acquired
acquires
acquiring
acquitted
adapted
add
added
adding
address
addressed
adds
adjust
adjusted
adjusts
administered
admires
admit
admits
admitted
adopt
adopted
ADOPTED
adopts
adorn
adorned
advanced
advancing
advertising
advise
advised
advises
advising
advocates
affect
affected
affecting
affects
affirmed
afford
afforded
aggravate
aggravated
aggravates
aided
aids
aimed
aiming
air
aired
airing
airs
alarmed
alleviate
allocate
allocated
allocating
allow
allowed
alter
altered
altering
amassed
amended
amplified
analyze
analyzed
anchored
angered
announce
announced
announcing
annoyed
answer
answered
answers
anticipated
anticipates
anticipating
appealed
appeased
applaud
applauded
apply
applying
appoint
appointed
appreciated
approach
approached
approaches
approaching
approve
approved
approving
arched
argued
argues
aroused
arrange
arranged
arrested
articulate
ascribe
ask
asked
asking
asks
assailed
assembled
assembling
assess
assessed
assessing
assign
assigned
assisting
assume
assumed
assumes
assured
astounds
attach
attached
attacked
attacking
attain
attended
attract
attracted
attracting
attracts
attribute
attributed
attributes
auction
auctioned
authorized
automates
averts
avoid
avoided
AVOIDED
avoiding
await
awaiting
awaits
awarded
backed
backing
backs
bankroll
banned
banning
bans
bar
barred
bars
base
based
bases
batter
battered
battle
battling
beat
beaten
beats
bedeviled
beefed
began
begin
begins
begun
believe
benefit
bequeathed
berated
besieged
bested
bestowed
bet
bid
bill
billed
bites
blame
blamed
blames
blew
block
blocked
blocking
blown
blunted
blurred
boast
boasts
bolster
bolstered
bolstering
bolsters
bolted
bomb
bombarded
booked
boost
boosted
boosting
boosts
bore
borrow
borrowed
bother
bothered
bottled
bought
bounces
boycotted
brace
braced
brandished
brands
braved
break
breaks
breathed
breathing
bred
breed
briefed
bring
bringing
brings
broke
brokered
brought
brushed
buck
bucked
buckling
buffer
build
building
builds
built
bumped
buoyed
burglarized
buried
burned
burnishing
buttressed
buy
buying
buys
calculate
calculated
call
called
calling
calls
calm
canceled
cap
capitalized
capped
captain
captioned
captured
carried
carries
carry
carrying
cart
carved
cast
casts
catapult
catch
categorized
caught
cause
caused
causes
causing
cautioned
ceased
ceded
ceding
celebrate
celebrated
celebrates
censored
chain
challenge
challenged
challenges

championed
change
changed
changes
changing
channel
chanted
characterized
charge
charged
charges
charging
chart
chartered
chase
chasing
chastised
chastises
check
checks
cheered
chew
chided
choose
choosing
chop
chose
chosen
chucked
cinch
circulated
circulating
cite
cited
cites
citing
claim
claimed
claims
clarifies
classified
classifies
cleans
clear
cleared
clears
clinched
clobbered
clocks
close
closed
clouds
co-anchored
cobbled
co-edits
co-founded
coined
collateralized
collect
collecting
color-coded
co-managing
combine
combined
combines
combing
command
commanded
commemorated
commissioned
committed
committing
compare
compared
compares
compensated
compiled
compiles
completed
complicated
compliment
composed
compressed
comprise
comprised
comprises
conceded
concedes
conceived
concentrated
concerned
concerns
conclude
concluded
concluding
concocted
condemn
condemned
condemns
conditioned
conduct
conducted
conducting
confers
confessing
confined
confirm
confirmed
confirms
confront
confronted
confronting
confronts
confused
confusing
conjures
consider
considered
considering
considers
consigns
consolidate
constrains
construct
constructed
constructing
consulted
consume
consumed
consummated
contemplate
content
contesting
continue
continued
continues
contradict
contradicts
contribute
contributed
control
controlled
controlling
controls
convene
convened
convert
converted
converting
convey
conveyed
convicted
CONVICTS
convinced
cool
coordinate
copy
cornered
correct
corrected
corroborate
couched
counseled
count
counted
counter
counts
courted
courting
courts
cover
covered
covers
crack
crafted
crams
cranked
create
created
creates
creating
credited
credits
criminalize
crippled
crisscrossing
criticize
criticized
criticizes
cropped
cross
crossed
crosses
crowded
crowds
crushes
cultivated
curbed
curtailed
curtailing
cushion
cushioned
cushioning
cut
cutting
damage
damaged
damp
damped
dampened
dangled
dashed
dated
dazzled
dealt
debate
debating
decide
decided
decimated
declared
declined
decontaminated
decrease
decreased
decreed
decried
decries
dedicated
deemed
defeat
defeated
defend
defended
defending
defends
defer
deferred
defied
defies
defined
defines
delay
delayed
DELAYS
deleted
delivered
delivering
delivers
demand
demanded
demanding
demolished
demonstrate
demonstrated
demonstrates
demonstrating
denied
denies
denounced
deny
depict
depicted
depicts
deplores
deployed
depress
depressed
derailed
derived
describe
described
describes
deserve
deserves
design
designed
designing
destroyed
destroying
destroys
detained
detected
deter
determine
determined
deterred
devastated
develop
developed
developing
develops
devise
devised
devote
devoted
devotes
diagnosed
did
diluted
diminish
direct
directs
disappoint
disappointed
disavowed
disbanding
discard
discarded
disciplined
disclose
disclosed
discloses
disconnected
discounted
discourage
discouraged
discourages
discover
discovered
discuss
discussed
discussing
disenchanted
dislike
dislikes
dismiss
dismissed
dismisses
disparage
dispatched
dispatching
dispelled
dispersed
display
displayed

displeases
dispute
disputed
disputes
disregarded
disrupt
dissipated
distinguished
distorts
distracted
distracting
distribute
distributed
distributes
distributing
disturbed
disturbs
divert
diverted
divided
divides
divorced
do
documenting
dodged
does
dogged
dogs
doing
doled
dominate
dominated
dominates
donated
donating
done
dons
dot
double
doubled
doubt
downed
downgraded
downplayed
drafted
drafting
drag
dragged
drain
drained
draw
drawing
drawn
draws
dreamed
drew
dried
drilled
drilling
drink
drive
driven
drives
driving
drop
dropped
dropping
drops
drove
drowned
dubbed
dug
dumped
dumping
dusted
dusting
dwarf
dwarfs
earmarked
earn
earned
earning
earns
ease
eased
eases
easing
eat
echo
echoed
eclipse
eclipsing
educated
eked
elected
electrified
elevated
elevates
eliminate
eliminated
eliminates
eliminating
eluded
emasculate
embargoed
embarrassed
embodied
embodies
embrace
embraced
embraces
embroiled
emphasize
emphasized
emphasizes
emphasizing
employ
employed
employs
empty
enacted
enclosed
encompass
encounter
encountered
encourage
encouraged
encourages
end
endangered
ended
ENDED
endorse
endorsed
ends
endure
endured
energized
enforced
engaged
engineered
engulfed
enhanced
enjoined
enjoy
enjoyed
enjoying
enjoys
enrich
ensconced
ensnarled
ensure
ensures
enter
entered
entering
enters
entertained
enticed
entombed
entrenched
envisaged
envisioned
equal
equals
erase
erased
erode
eroded
escaped
eschewed
espouse
establish
established
establishes
estimate
estimated
estimates
evaluate
evaluates
evidenced
evinced
evolve
evolved
exacerbate
exacerbated
exacerbates
examined
exceed
exceeded
exceeding
exceeds
exchange
exchanged
excised
excited
exclude
excluded
excoriated
execute
executed
executes
executing
exemplifies
exempted
exercise
exercised
exercises
exhausted
exited
expand
expanded
expanding
expands
EXPANDS
expect
expected
expecting
expects
expedite
expelled
experience
experienced
experiencing
explain
exploit
exploring
export
exported
exporting
exports
expose
exposed
expressed
expresses
extend
extended
extending
extends
extinguish
extorted
extract
extrapolated
exuded
eyed
eyeing
eying
face
faced
faces
facing
FACING
factors
faked
fanned
farms
fashioned
favor
favored
favors
fear
feared
fears
feature
featured
features
fed
feed
feeling
fell
felt
fended
fertilized
fetch
fetched
fetches
fielded
fight
fighting
figure
file
filed
fill
filled
filling
fills
filmed
finagled
finalized
finance
financed
financing
find
finding
finds
fined
finessed
fingered
finish
finished
fired
fires
fit
fits
fixed
flash
flashed
flashes
flattened
flaunt
flawed
fled
floated
flooded
flunk
flunking
fly
focused
focusing
foiled
fold
follow
followed
following
follows
force
forced
forces
forcing
forecast
forecasting
forecasts
foresaw
foresee
foreseen
foresees

forget	grabbed	helping	impress	interrupted	knows
forgotten	granted	helps	impressed	intersperses	laced
form	grants	herald	imprisoned	interviewed	lack
formed	grasp	heralded	improve	intimidating	lacked
forming	greeted	hidden	improved	introduce	lacks
forsaken	grounded	hide	improves	introduced	laid
fostered	grouped	highlight	improving	introduces	lambasted
fought	grown	highlighted	inaugurated	introducing	lamented
fouled	grows	highlights	include	inundated	lauded
found	guarantee	hinder	included	invade	launch
founded	guaranteed	hindered	includes	invaded	launched
fractured	guarantees	hired	incorporated	invent	launches
fragmented	guide	hires	incorporates	invented	launching
frayed	guided	hiring	increase	inverted	lavishing
freed	gunned	hit	increased	invest	lay
FREED	hailed	hitting	increases	invested	laying
frees	halted	hobbled	increasing	investigated	lays
freezes	halts	hoisted	incur	investigates	lead
frequents	halved	hold	incurred	investigating	leading
frighten	hammered	holding	indicate	invests	leads
frightened	hampered	holds	indicated	invite	learn
frittered	hampering	honed	indicates	invited	learned
froze	hampers	honor	indicted	invites	learning
frozen	hamstring	hoped	inferred	invoke	learns
frustrated	hamstrung	hospitalized	infiltrating	irked	lease
fueled	hand	hosting	inflated	irritates	leased
funded	handed	hosts	inflates	isolated	leases
funding	handle	houses	influence	issue	leave
funds	handled	humbled	influenced	issued	leaves
funneling	handles	hurt	influences	issues	leaving
gain	handling	hurting	influencing	jacked	lectures
gained	hands	hurts	inform	jailed	led
gaining	hanging	hypnotized	informed	jammed	left
gains	harbor	identified	informs	join	lend
galvanized	harbors	identifies	infuse	joined	lends
gathered	hastened	identify	infused	joining	lent
gauges	hate	idle	inherited	joins	let
gave	hauled	idled	inhibited	jolt	leveled
generate	haunt	ignore	initialed	jolted	levied
generated	haunting	ignored	initiate	junk	liberalized
generates	head	ignores	initiated	jury-rigged	liberated
get	headed	illustrate	injected	justified	license
gets	heads	illustrated	innovated	justifies	licensed
getting	heaped	illustrates	inserted	kayoed	licenses
give	hear	imagine	inspect	keep	lifted
given	heard	impacted	inspire	keeping	like
gives	hearing	impeded	inspired	keeps	liked
giving	hears	implanted	install	kept	likened
gleaned	heartened	implemented	installed	kicked	likes
glutted	heaved	implicated	instigated	kill	limit
gobbled	heckled	implies	instituted	killed	limited
got	hedge	imported	insures	killing	limiting
gotten	heightened	impose	intensified	kills	limits
govern	held	imposed	intercepted	knocked	lined
governed	help	imposes	interpret	know	linked
grab	helped	imposing	interpreted	known	liquidated

list
listed
lists
lit
loaded
loathed
lobbied
located
locked
lodged
logged
loosen
lorded
lose
loses
losing
lost
love
loved
lower
lowered
lowering
lured
lures
made
magnified
mailed
maintain
maintained
maintaining
maintains
make
makes
making
MAKING
manage
managed
manages
managing
mandate
maneuvered
manipulates
manipulating
manned
manufactures
manufacturing
map
mapped
mark
marked
market
marketed
marketing
markets
marks
marred
married
mastered
match
matched
matches
matching
mated
mean
means
meant
measure
measured
measures
meet
meets
melds
melt
memorize
mention
mentioned
mentions
merged
merit
met
metabolized
mimic
minimize
minted
mired
miscalculated
miss
missed
misses
misstated
misstates
mistaken
mocked
modeled
modernized
modifies
modify
molded
monitor
monitored
monitoring
mop
mortgage
motivated
motivating
mounted
mounting
mounts
move
moved
moves
moving
muddied
muster
muted
name
named
narrowed
narrows
need
needed
needs
neglected
neglecting
negotiated
negotiating
net
netted
nicknamed
nods
nominated
normalize
notched
note
noted
notice
notified
notify
noting
nullified
numbered
obey
observed
obtain
obtained
occupy
offend
offer
offered
offering
offers
offset
offsetting
ogling
omit
omits
omitted
open
opened
opening
opens
operate
operated
operates
operating
oppose
opposed
opposes
order
ordered
organized
organizing
ousted
outdid
outgained
outlawed
outleaped
outlined
outlines
outnumbered
outpaced
outpacing
outperformed
outraged
outselling
outsells
outshines
outstripped
outstrips
outweighed
overcame
overcome
overhauled
overlays
overlooked
overlooks
over-magazined
overrode
oversaw
oversee
overseeing
overshadowing
overstated
oversubscribed
overturned
overweighted
overwhelmed
owe
owed
owes
own
owned
owns
pack
packed
packs
paid
paint
painted
paired
parachute
parallel
parallels
paralyzed
parcel
pardoned
pare
pared
parked
pass
passed
passing
pasteurized
patented
paved
pay
paying
pays
peddle
peddles
peg
pegged
penalize
penalized
penetrate
penetrated
perceived
perform
performed
performing
permit
permits
permitted
perpetuates
persuaded
persuades
phase
phasing
phoned
pick
picked
picks
pictures
piled
pilloried
pinpointed
pin-pointed
pioneered
piped
pitched
pitches
pitching
pits
pitted
place
placed
places
placing
plagued
planned
planning
plans
plant
planted
play
played
playing
plays
Plays
pleaded
please
pleased
pledged
plotting
plow
plows
plugged
plunge
plunged
plunking
point
pointed
pointing
points
poked
police
polled
pondering
pooled
popularized
portends
portray
portrayed
pose
position
possess
possesses
post
posted
posting
postpone
postponed
pour
poured
pouring
powered
practicing
praise
praised
praises
preach
preapproved
preceded
precipitated
precluded
precludes

predicated
predict
predicted
predicting
predicts
prefer
preferred
prefers
premiere
preoccupied
prepared
preparing
pre-register
prescribe
prescribed
prescribes
presented
presents
preserved
pressed
pressing
pressure
pressured
prevent
prevented
prevents
previewing
priced
print
printed
privatize
process
processed
processes
produce
produced
produces
producing
professes
proffered
prohibit
prohibited
prohibits
project
projected
projecting
prolong
promise
promised
promises
promising
promote
promoted
promotes
promoting
prompt
prompted
pronounced
propelled
propelling
proposed
proposes
propped
protect
protected
protecting
protects
protested
protests
prove
proved
proven
provide
provided
provides
providing
provoked
pruned
publish
published
publishes
publishing
pull
pulled
pulling
pulls
pummeled
pumped
pumping
purchase
purchased
purchases
pursued
pursuing
pushed
pushing
put
puts
putting
puzzled
quashed
question
questioned
questioning
questions
quieted
quoted
quotes
racked
racking
raided
rained
raise
raised
raises
raising
rallied
ran
rang
ranked
ranks
rated
ratified
rattle
rattled
RATTLED
reach
reached
reaching
reactivated
read
readied
reading
reads
reaffirmed
realized
realizes
realizing
REAP
reaped
reaping
rearranges
reasserting
reasserts
reassigned
reassured
rebuffed
rebuilding
rebuked
recall
recalled
recalls
recanted
receive
received
receives
receiving
recentralized
recites
reclaiming
reclaims
reclassified
recognize
recognized
recognizes
recommend
recommended
recommends
reconcile
reconnect
reconstructed
recorded
records
recounted
recounts
recoup
recouped
recover
recovered
recovering
recruited
redeemed
redesigned
redirected
redoubling
redraw
reduce
reduced
reduces
re-elected
reeled
re-enacting
refers
refinanced
refining
reflect
reflected
reflecting
reflects
refund
refused
refuted
regained
regard
regarded
regards
registered
regulated
reignited
reimburse
reinforce
reinforced
reinforcing
reinstated
reintroduced
reinvented
reinvigorate
reiterated
reiterates
reject
rejected
rejoined
related
relaunched
relaxed
relayed
release
released
releases
relegated
relieve
relieved
relinquished
relished
remember
remembered
remembers
reminded
reminds
remodeling
removed
removes
renegotiated
renewed
renewing
renews
renounced
rent
reoffered
reopened
reoriented
repackaged
repaid
repainted
repaired
repay
repeals
repeat
repeated
repel
replace
replaced
replaces
replacing
replenished
replied
report
reported
reporting
reports
represent
represented
representing
represents
repriced
reprinted
repudiate
repurchased
requested
requesting
require
required
requires
requisitioned
rescheduled
rescinded
rescued
researched
resemble
resembles
resent
reserving
reset
reshaped
reshaping
resigned
resist
resisted
resisting
resolve
resolved
rest
restated
restore
restored
restrict
restricted
restricts
restructured
resume
resumed
resurrect
resurrected
retain
retained
retains
rethinking
retraced
retracted
return
returned
returns
reunited
revamp
revamped
reveal
revealing
reversed
reverses
review
reviewed
reviewing
reviews
revised
revising
revive

revived
revoke
revolutionized
revved
reward
rewarded
rewrite
ride
ridicules
rigged
ripped
risk
riveted
robbed
rocked
rode
roil
roll
rolled
rolling
rolls
romancing
round
rounded
routed
routes
rub
rubbed
rubs
rule
ruled
run
rung
running
runs
rushed
sacked
sacrifices
saddled
said
salted
saluted
salvaged
sanctioned
sandwiched
sang
saturated
save
saved
saves
savored
saw
say
saying
says
scaled
scandalized
scanning
scans
scare
scared
scheduled
scolded
scooped
scoops
scored
scotched
scouring
scouting
scrapped
scraps
scratch
scrutinized
scrutinizing
scuttle
scuttled
searched
seated
secured
see
seeing
seek
seeking
seeks
seen
sees
seize
seized
select
selected
selects
sell
selling
sells
send
sending
sends
sense
sent
sentenced
separate
sequester
serve
served
serves
services
set
sets
setting
settle
settled
settling
severed
shags
shake
shaken
shakes
share
shared
sharpening
shattered
shed
shielded
shift
shifted
shifts
shipping
shocked
shook
shoot
shot
shouldering
shouting
shouts
shove
show
showed
showing
shown
shows
shredded
shrinking
shrouded
shrugged
shun
shunning
shut
shuts
sidestepped
sign
signal
signals
signed
signing
signs
simulate
single
singled
sinking
siphoned
siphoning
sipped
sketches
skewed
skipped
skipping
skirted
slammed
slapped
slaps
slashed
slated
slimmed
slow
slowed
slowing
smoked
snapped
snaps
snatched
sneak
sniffing
snorts
snubbing
softened
soiled
sold
soliciting
solidified
solidify
solve
solved
sought
sounded
sour
sow
sowed
spared
spark
sparked
sparking
spawned
speak
spearheaded
specified
specify
speed
speeds
spell
spells
spend
spending
spends
spent
spies
spins
splashed
split
splits
sponsor
sponsored
sponsoring
spook
spooked
sports
spotted
spread
sprinkle
sprinkled
spun
spur
spurn
spurned
spurred
spurring
spurs
squandered
squeezed
squelched
staffed
staged
stalking
stalled
start
started
starting
starved
stashed
stated
stave
steal
step
stepped
stepping
steps
stick
stimulated
stir
stirred
stirring
stock
stocks
stoked
stolen
stop
stopped
stored
strained
streamlined
strengthen
strengthened
strengthens
stressed
stresses
stretching
stricken
strike
strikes
stripped
struck
STRUCK
structured
stuck
studded
studied
studying
stuffed
stunned
subdued
subjected
submit
submitted
submitting
subpoenaed
subsidize
subsidized
substitute
subverted
succeed
succeeded
succeeds
sued
suffer
suffered
suffering
suffers
suggest
suggested
suggesting
suggests
suing
summarize
summed
summoned
sung
supervising
supplement
supplied
supplies
supply
supplying
support
supported
supporting
supports
suppressed
surprise
surprised
surrendered
surrounded
survey
surveyed
surveys
survive
survived
survives

suspects
suspended
suspending
sustained
sustains
swallowed
swamped
swap
swayed
sweeping
sweeps
sweetened
swell
swelled
swept
switch
switched
symbolized
symbolizes
tacked
tackle
take
taken
takes
taking
tallied
tallying
tangled
tap
taped
tapped
target
targeted
targeting
targets
tarnish
tarnished
taught
taxed
teaches
telegraph
tell
telling
tells
termed
terminated
test
tested
testing
thought
threaten
threatened
threatens
threw
thrill
throw
thrown
throws
thrusting
thwart
thwarted
ticks
tied
tighten
tightened
tightening
timed
told
toned
took
top
topped
toppled
tops
torched
torments
torn
torpedo
torpedoed
tossed
totaled
tote
toted
touched
tours
touted
touting
touts
traced
track
tracks
trade
traded
trades
trailed
trails
transfer
transferred
transferring
transfers
transformed
transforms
translated
transported
transports
traumatized
travel
treat
treated
treats
tried
triggered
trim
trimmed
trimming
tripled
tripped
trotted
troubled
trust
try
trying
turn
turned
turning
turns
twist
twisted
twists
typed
typified
typifies
uncovered
undercut
undercutting
underestimated
underfunded
undergoing
underlined
undermine
undermined
underperform
underscore
underscored
underscores
understand
understands
understate
undertaken
undertook
underwhelmed
underwrite
underwritten
undone
unfazed
unhinged
unleashed
unload
unlocked
unmasks
unnerved
unravel
unroll
unveil
unveiled
unveiling
upgraded
upgrading
upheld
upset
urged
urges
urging
use
Use
used
uses
ushered
ushers
using
vacate
value
valued
values
vary
vented
ventilated
veto
vetoed
view
viewed
views
violate
violates
visit
visited
visiting
visits
voiced
voices
volunteered
voted
vowed
waged
waging
waived
waiving
warned
warns
washed
waste
wasted
wastes
watch
watched
watches
watching
waved
waving
weakened
wear
wears
weather
weds
weigh
weighed
weighs
weighted
welcome
welcomed
welcomes
whipped
whipsaw
whittled
widen
wields
win
winning
wins
wipe
wiped
wish
wishing
withdrawn
withdrew
withheld
witnessing
woken
won
WON
wooing
wore
work
worked
working
works
worried
worries
worsened
wrack
wrapped
wraps
wreak
wrecked
wrestles
write
writes
writing
written
wrote
yanked
yelped
yield
yielded
yielding
yield

Chapter 3

The Person Hierarchy

While the precise formulation of the person hierarchy has been debated, some hierarchy of person has proved useful in accounting for phenomena including split ergativity, the distribution of direct and inverse verbs in inverse languages, and the distribution of null and overt subjects. Most interestingly and relevantly here, the hierarchy of person has been invoked to account for the distribution of active and passive in a number of languages, with local (first and second) person patients driving passivization and local person agents frequently suppressing it. In this chapter I will present data supporting the hypothesis that the person hierarchy has an effect on the frequency of passivization in English. I will begin by reviewing arguments and debate over the precise form the hierarchy should take and ways in which its effects on syntax have been accounted for. Then I will review the relevant data in languages that show interactions of person and voice and discuss previous arguments and data regarding the effects of person on voice in English. Finally, evidence will be presented supporting the hypothesis that the person hierarchy has an effect on the frequency of passivization in English. Parallel to the previous chapter, this effect will be modeled using the constraints produced by harmonic alignment of the person hierarchy with the grammatical relations hierarchy (as discussed in Aissen (1999)) and the stochastic optimality theory framework of Boersma and

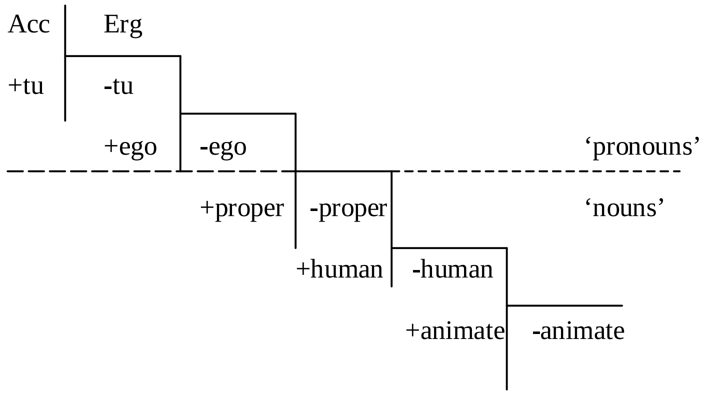
Hayes (2001). The material to be discussed here is largely parallel to Bresnan, Dingare, and Manning (2001); the difference is mainly in greater discussion of the person hierarchy, more detailed description of the methodology used in the corpus work, and some additional data regarding the relative ranking of local persons in English.

3.1.2 Grounding

Below I will review several approaches to the grounding of the person hierarchy – that is, approaches seeking to explain why different elements of the hierarchy are treated differently in certain phenomena. These approaches tend to fall into different classes in terms of how they characterize elements higher on the hierarchy. One approach associates these elements with the likelihood of being agents – this approach attempts to explain facts of split ergativity. Another class of approaches associates elements higher on the hierarchy with some form of greater accessibility to the speaker. The approaches also tend to diverge in the elements they rank; what will be addressed as "the person hierarchy" here has often been addressed as part of a larger animacy hierarchy, or in the case of Silverstein (1976), a hierarchy involving person, definiteness, and animacy. Here, I will only consider the specific dimension of "person" comprising the elements of first, second and third person, and the approaches below are of interest in how they rank those elements.

Greenberg (1966) may have been the first to propose a markedness scale for person – his scale, based on facts from verb agreement and simple frequency counts of the occurrence of first, second, and third person pronouns, took the form 3 > 1 > 2. However, what is generally referred to as the person hierarchy takes the opposite form, ranking local persons above third persons, and is most associated with Silverstein (1976). Silverstein (1976) does not take the approach of claiming certain persons to be universally more marked than others, but of deeming specific associations of person and semantic role as marked or unmarked. His claim is advanced to account for facts of split ergativity; in his hierarchy, reproduced below, elements at the top are least likely to be marked with ergative case marking, while positions at the bottom are least likely to be marked with accusative case marking. Silverstein therefore suggests that elements at the top of the hierarchy are the most natural (or unmarked) agents of true transitive verbs, while elements at the bottom are the most natural patients.

(1)



The approach of associating elements higher on the hierarchy with a higher likelihood of being agents is also taken by Dixon (1979), who argues from intuition that the speaker sees himself as the "quintessential agent", the person with whom he is speaking less so, and so on down the hierarchy. This is argued against by Wierzbicka (1981), who argues that the first person sees himself as the quintessential experiencer, not the quintessential agent, and by DeLancey (1981), who argues that while the notion of natural agency may explain a ranking of animate > inanimate, it does not particularly explain the ranking of local persons above third person humans, which is by far the most common pattern of split ergativity, while the pattern animate > inanimate is relatively rare.

Givón instead grounds the person hierarchy in topicality, arguing that local persons are statistically more likely to be the topic of unmarked clauses than third persons (Givón 1994), and that first and second persons are the most presupposed arguments in the discourse (Givón 1976). Along similar lines, Kuno and Kaburaki (1977) and Kuno (1987) ground the person hierarchy (their "speech act hierarchy") in the concept of *empathy*, where empathy is defined as the speaker's identification with a person who participates in the event that he describes in a sentence. Kuno and Kaburaki argue that the speaker empathizes the most with himself and then with the addressee, and the least with third persons. Finally, within the accessibility theory framework discussed in the previous chapter, Ariel (2000) shows that local person referents are referred to with high accessibility markers at a significantly greater frequency than third persons. Since passivization has been characterized as being driven by topicality, givenness, and empathy status of the patient argument, and plausibly could be linked to its accessibility, these approaches clearly motivate increased passivization with local person patients and decreased passivization with local person agents.

The harmonic alignment approach of Aissen (1999) to be adopted here lies between these two approaches; it simply takes the person hierarchy as a prominence scale (which could be associated with empathy, topicality, accessibility, and so on), with higher elements tending to be subjects and lower elements non-subjects. Her approach yields similar predictions about voice and person as above. It should be noted, however, that Aissen herself analyzes these facts as the expression of marked associations of person and semantic role in the marked voice (passive). Markedness of passives appears to be important to her argument only for this reason.

3.1.2 Relative Ranking of First and Second

The order of the first and second person elements on the person hierarchy has varied considerably from account to account. Silverstein's hierarchy spans features from several dimensions, including animacy, definiteness, and person, but as can be seen, the person hierarchy implied by (1) is 2 > 1 > 3. The speech act empathy hierarchy of Kuno and Kaburaki (1977) takes the form "Speaker > Hearer > Third Person" (1 > 2 > 3), but Kuno (1987) revises this to "Speaker > Others" (1 > 2 & 3), stating that whether the speaker feels greater empathy with the hearer or with third parties cannot be predetermined. Data reviewed in Ariel (2000) does not point to any consistent difference in accessibility of first and second persons; a journalistic text examined (Levy 1995, as cited in Ariel 2000) shows second persons to be much more accessible than first persons, while data from Hebrew conversations (Lotan 1990, as cited in Ariel 2000) shows first and second persons to be equally accessible, and data from interviews and short stories show first persons to be more accessible than second persons. Finally, DeLancey (1981) cites distribution of the inverse marker as evidence that Nocte attests the hierarchy 1 > 2 > 3 while Algonquian attests the hierarchy 2 > 1 > 3. Thus, it appears that the relation between first and second person may be indeterminate, or may vary from genre to genre, or from language to language. Therefore DeLancey (1981) and Aissen (1999) take the position that local persons are universally ranked higher than third persons, but the ranking within local persons (i.e., between first and second persons) is language-particular.

3.2 Formalizing the Effects of Person On Voice

To formalize person hierarchy effects, Aissen uses the technique of harmonic alignment as in the previous chapter to associate the person and semantic role hierarchies with particular grammatical relations, and to generate optimality theoretic constraint hierarchies penalizing marked associations.

The binary scale of Su > Non-Su is aligned with the person hierarchy. These two scales are shown below:

(2) Grammatical Relations Hierarchy: Su > Non-Su
(3) Person Hierarchy: Local > 3

Harmonic alignment of the hierarchy in (2) with the hierarchy in (3) yields the following harmony scales and corresponding constraint subhierarchies:

(4) Su/Local ≻ Su/3 *Su/3 >> *Su/Local
(5) Non-Su/3 ≻ Non-Su /Local * Non-Su /Local >> * Non-Su /3

Replacing "Non-Su" in (6) with "Object" and "Oblique" yields (7) and (8):

(6) Object/3 ≻ Object/Local *Object/Local >> *Object/3
(7) Oblique/3 ≻ Oblique/Local *Oblique/Local >> *Oblique/3

In languages treating first and second persons differently, the constraints involving local persons can be separated into separate constraints addressing only first and only second persons (for example, *Object/Local can be divided into *Obj/1st and *Obj/2nd).

The implications of this constraint set are parallel to the implications of the definiteness constraints. Again, it is problematic that the constraints do not imply that passivization will occur only when the patient is higher on the hierarchy than the agent. For example, in Aissen's framework one could have obligatory passivization with third-person agents and third-person patients by a high ranking of the constraint *Object/3. If one extended the person hierarchy to the animacy hierarchy, one could have obligatory passivization for first person agents and animate (or higher) patients but not first-person agents and inanimate patients. The rankings producing this scenario are reproduced below.

Obligatory Passive

/Local Agent – Local Patient/	*OBJ/LOCAL	*OBJ/ANIM	*OBLIQUE/LOCAL	*OBJ/INANIMATE
☞ [Passive]			*	
*[Active]	*!			

Obligatory Passive

/Local Agent – Animate Patient/	*OBJ/LOCAL	*OBJ/ANIM	*OBLIQUE/LOCAL	*OBJ/INANIMATE
☞ [Passive]			*	
*[Active]		*!		

Obligatory Active

/Local Agent – Inanimate Patient/	*OBJ/LOCAL	*OBJ/ANIMATE	*OBLIQUE/LOCAL	*OBJ/INANIMATE
☞ [Active]				*
*[Passive]			*!	

In these tableaux, the subhierarchy *Obj/Local >> *Obj/Animate >> *Obj/Inanimate is respected, but we can still generate a language in which passive is obligatory for a patient lower on the animacy hierarchy than the agent (in the second tableau), even though the language does have active. This seems particularly problematic in the case of person because the observed cases of person-voice or animacy-voice interaction point to, and have been analyzed as conforming to, the generalization that the subject must not be lower on the hierarchy than the object (e.g. Klokeid 1978 (as cited in Whistler 1985), Chung 1998). Thus, as stated in the previous chapter, this constraint set seems more suited to modeling subject and object marking (where marking of the object is independent of the status of the subject) than to modeling passive. Again, for our purposes these problems will simply be noted. As will be seen, the constraint set suffices to model frequencies of passivization.

By the same reasoning as in the previous chapter, it holds that the constraints in (1)-(6) imply, (disregarding other constraints), that if a language has obligatory passivization for inputs with agents of person status X and patients of person status Y, then it will have obligatory passivization for inputs with agents of person status X and patients of person status Z > X, and for inputs with agents of person status Z < X and patients of person status Y. That is, in the table below, if there is obligatory passivization at any square, then there must be obligatory passivization for all squares to the left and for all squares below. In Boersma's framework, this translates into the prediction that (again disregarding other constraints) if passivization occurs at a particular frequency in some square, it must occur at an equal or lower frequency for all squares to the right and at an equal or greater frequency for all squares below.

Table I. All Possible Inputs Specified for Person and Role

Agent ↓ Patient→	Local person	Third person
Local person		
Third person		

In the next section I will review evidence supporting these predictions cross-linguistically and in English.

3.3 Cross-Linguistic Person-Voice Interactions

3.3.1 Categorical Interactions

A small number of languages exhibit categorical person-voice interactions (see Bresnan, Dingare, and Manning 2001 for a list; part of the data discussed here is from the languages and works cited in that list). That is, certain configurations of person in agent and patient are obligatorily expressed in the active or obligatorily expressed in the passive. Here, I briefly review some of these interactions and how they would be modeled by Aissen's constraints. When the agent is a local person, passivization is prohibited in the Coast Salish languages Lummi, Squamish, and Lushootseed, (Jelinek and Demers 1983), as well as in the Nootkan languages (Nootka, Nitinat, and Makah) of Vancouver Island and Washington State (Whistler 1985), and in Southern Tiwa (Allen and Frantz 1978) and Picurís (Zaharlick 1982), languages of New Mexico. This can be modeled by a high ranking of *Oblique/Local. The second common restriction in languages with person-voice interactions is a prohibition against local person objects. Some prohibition on actives with local person patients (often when the agent is lower on the hierarchy) occurs in Lummi, Squamish (only for second person), Bella Coola, the Nootkan languages, Southern Tiwa, Picurís, Arizona Tewa (Kroskrity 1985), and Chamorro (Chung 1998). In these languages, the passive must occur instead, so that *Obj/Local must be ranked over *Su/Pat as depicted below:

Table II.

/Local Person Patient/	*OBJ/LOCAL	*SU/PAT
☞ [Passive]		*
*[Active]	*!	

Languages which have prohibitions against passivization with local person agents and prohibitions against local person objects will have different outputs when both agent and patient are local person depending on the relative ranking of *Obj/Local and *Oblique/Local. Languages which rank *Oblique/Local higher than *Obj/Local will prohibit passivization when both agent and patient are local person. This appears to be the case in all the languages except Arizona Tewa (for Chamorro and Bella Coola no data could be obtained).

Table III.

/Local Agent – Local Patient/	*OBLIQUE/LOCAL	*OBJ/LOCAL	*SU/PAT
☞ [Active]		*	
*[Passive]	*!		*

In Arizona Tewa, the opposite holds. Active sentences obligatorily encode third person objects, so that first person patients, even with local person agents, are obligatorily passive. (Arizona Tewa appears to be a counterexample to the generalization that passive will only be obligatory when the patient is higher in person than the agent.) So *Oblique/Local must be ranked lower than *Obj/Local:

Table IV.

/Local Agent – Local Patient/	*OBJ/LOCAL	*OBLIQUE/LOCAL	*SU/PAT
☞ [Passive]		*	*
*[Active]	*!		

Overall, it seems that it is most common for *Oblique/Local to outrank *Obj/Local, although as stated previously no data could be obtained for Chamorro and Bella Coola.

3.3.2 Effects of Person on Voice in English: Preferences and Frequencies

The effects of person on the acceptability of active and passive sentences in English is examined most closely by Kuno and Kaburaki (1977) and Kuno (1982). As stated previously, their account is driven by a concept of *empathy*, where empathy is defined as "the speaker's identification, *with varying degrees* (ranging from degree 0 to 1) with a person who participates in the event that he describes in a sentence." In their account, certain sentence structures imply certain empathy relationships, and these relationships must be consistent with one another and with certain universal empathy hierarchies. This is formalized with "The Ban on Conflicting Empathy Foci" which states that "a single sentence cannot contain logical conflicts in empathy relationships". Because of their "Surface Structure Empathy Hierarchy" which states that subjects are easier to empathize with than non-subjects, active sentences entail that the speaker has greater empathy with the agent than the patient, and passives entail that the speaker has greater empathy with the patient than the agent. These relationships must be consistent with the "Speech Act Participant Empathy Hierarchy" which states that the speaker always feels the greatest empathy for himself (or for himself and then the addressee in Kuno and Kaburaki 1977).

Thus, the account predicts acceptability for sentences such as *I met Mary* and *I was hit by Mary* since the surface structure empathy hierarchy implies that the speaker's empathy with the subject (himself) is greater than his empathy with the object or oblique (*Mary*), and this is consistent with the Speech-Act Participant Empathy Hierarchy; therefore the sentence is acceptable. The sentences *Mary was hit by me* or *Mary hit me*, on the other hand, imply through the surface structure empathy hierarchy that the speaker's empathy with Mary is greater than his empathy with himself, which contradicts the Speech-Act Participant Empathy Hierarchy. Thus these sentences violate The Ban on Conflicting Empathy Foci and should be unacceptable. Kuno (1982) formally accounts for the acceptability of *Mary hit me* with his "Markedness Principle for Discourse-Rule Violations", which states that: "Sentences that involve marked (or intentional) violations of discourse principles are unacceptable. On the other hand, sentences that involve unmarked (or unintentional) violation of discourse principles are acceptable." That is, since the active is unmarked, it is always acceptable even when it violates the ban on conflicting empathy foci. However, since the passive is marked, it is unacceptable when it violates the ban on conflicting empathy foci.

Kuno and Kaburaki are unclear on whether they intend their account to predict ungrammaticality or simply dispreferredness or marginality; however, Kato (1979) argues in response that there is no constraint against first-persons as logical subjects in English passives given the appropriate discourse context. In support of this he gives examples including the following:

(1) I said, "Me watch it! Fuck that! Let him watch it." He was hired by me. I could fire him if I didn't like him. (Studs Terkel, *Working*)
(2) When somebody says to me, "You're great, how come you're *just* a waitress?" *Just* a waitress. I'd say, "Why, don't you think you deserve to be served by me?" (Studs Terkel, *Working*)
(3) Gore [Vidal] never lacked love, nor was he abandoned by me. (*Time*)
(4) The slight silken scrape of her knock-knees when she walked quickly was, I repeat, highly prized by me. (Saul Bellow, *Humboldt's Gift*)
(5) "If you would be guided by me – ", he said, hesitating. (Agatha Christie, *The Man in the Brown Suit*)
(6) It was the same with my eighth period class who were told by me to either come watch or get into the film or just stay in the room. (James Herndon, *How to Survive in Your Native Land*).

Kato's examples demonstrate that local person passive agents are grammatical in English given the appropriate discourse context. The account of Kuno and Kaburaki, however, still suggests that the constraints on person introduced in the last section could reveal themselves through low frequencies of marked person-voice configurations, thus confirming the predictions of a stochastic optimality theory interpretation of these constraints. Below I briefly examine some preliminary data supporting this speculation.

The data of Svartvik (1966) (reviewed in the previous chapter) demonstrate that pronoun agents are less likely to passivize and pronoun patients more likely to passivize. While Svartvik does not provide separate data for local person versus third person pronouns, his results would be consistent with a lower likelihood of passivization with local person agents and a higher likelihood for local person patients. Estival and Myhill (1988) show that local person patients are more likely to passivize (29%) than third-person pronoun patients (12%) and nominal patients (5%); their data is consistent with the predictions outlined above but does not address agents. Elena Seoane Posse (2000) conducts a corpus study of passive in the early modern English period examining data concerning semantic role, animacy, and person. It is difficult to draw conclusions from her study with regard to the person data since she only looks at passive clauses, and of her 349 passive clauses only 14 contain speech act participants. Of these, 12/14 were the subjects rather than the *by*-phrase of passive clauses. However, since she does not provide corresponding data for actives, this data is difficult to interpret and could be the result of a higher frequency of pronominal patients. Also, the numbers are quite small, making it difficult to draw reliable conclusions. In the next section, I present a more detailed study examining the specific frequency-gradation predictions for the four possible inputs in a corpus of English.

3.4 A Corpus Study Of The Effect of Person on Frequencies of Passivization in English

3.4.1 Methodology

The methodology was for the most part parallel to the study of definiteness. Rather than the Wall Street Journal corpus, the Switchboard, another sub-corpus of the Penn Treebank, was used. The Switchboard Corpus consists of approximately 2400 telephone conversations and is parsed and annotated. The primary reason for choosing this corpus was the high frequency of local pronouns (as is shown in Francis et al (1999), 91% of subjects in the Switchboard are pronominal). Like the Wall Street Journal Corpus, the Switchboard Corpus can be easily searched using the tgrep program, which allows the user to specify patterns for the tree structure of a sentence, and then returns all the trees in the corpus corresponding to that pattern. The goal was to find the numbers of active and passive outputs in the corpus corresponding to the following four inputs:

1. /local person agent + third person patient/
2. /local person agent + local person patient/
3. /third person agent + local person patient/
4. /local person agent + third person patient/

For example, for the first input we would search for the number of active sentences with local person subjects and third person objects and the number of passive sentences with third person subjects and local persons in the oblique. As in the definiteness study, only full *by*-phrase passives were counted, since it would be difficult to determine which of inputs (1)-(4) an agentless passive corresponded to (due to the absence of the agent argument). However, figures for agentless passives were obtained as supplementary data.

In addition to the inputs considered above, a separate count was also done in which "local person" was separated into first person and second person. That is, searches were completed to find the numbers of active and passive outputs in the corpus corresponding to the following nine inputs:

1. /first person agent + first person patient/
2. /first person agent + second person patient/
3. /first person agent + third person patient/
4. /second person agent + first person patient/
5. /second person agent + second person patient/
6. /second person agent + third person patient/
7. /third person agent + first person patient/
8. /third person agent + second person patient/
9. /third person agent + third person patient/

The purpose of doing this was to test whether a subranking of first and second persons in English could be detected. This ranking clearly varies cross-linguistically, and its status in English has been unclear (cf. section 3.1.2).

For the same reasons as described in the definiteness study, only main verbs were considered, sentences containing empty subjects (as in imperatives) were removed, sentences containing coordinated subjects, logical subjects, and objects were removed, and a list of nonpassivizing verbs was compiled and those verbs were excluded. Actives, passives, and topicalizations were detected in the same way as in the definiteness study (except that, in searching for objects, we searched up to the tenth daughter), so the scripts are not reproduced here. The lists of verbs judged to be passivizing and non-passivizing in the majority of their senses appear in the appendix.

Local person noun phrases were detected as those noun phrases whose leftmost daughter was a local pronoun, and all noun phrases which were not local noun phrases and were not empty or expletives were classified as third person noun phrases. Similarly, first-person noun phrases were detected as those noun phrases whose leftmost daughter was a first-person pronoun, and similarly for second-person noun phrases. In this sense our definition of local and third person noun phrases was syntactic. That is, the noun phrase *the author,* while it can be used by an author to refer to himself, is still syntactically third person and would be counted as such in our study. The tgrep commands defining first-person, second-person, local-person (that is, first and second combined for the first study), and third-person noun phrases are reproduced in the appendix.

3.4.2 Results & Analysis: Local vs. Third Person

The raw data is shown in the table below:

Table V. Raw Data

Agent ↓ Patient→	Local Person	Third Person
Local Person	Actives: 179 (179 +0 topicalized) Passives: 0 (0 +0 topicalized)	Actives: 6246 (6217 +29 topicalized) Passives: 0 (0 +0 topicalized)
Third Person	Actives: 472 (472 +0 topicalized) Passives: 14 (14 +0 topicalized)	Actives: 3110 (3107 +3 topicalized) Passives: 39 (39 +0 topicalized)

The rate of passivization for each input was calculated as the number of passives divided by the number (actives plus passives) and is presented in the table below.

Table VI. Rate of Passivization

Agent ↓ Patient→	Local Person	Third Person
Local Person	0%	0%
Third Person	2.9%	1.2%

Table VIII. Significant Differences

A ↓ P→	Local Person	Third Person
Local Person	0%	0%
Third Person	2.9%	1.2%

These results are in line with the hypothesis that the rate of passivization decreases or remains the same from left-to-right in rows and increases or remains the same from top-to-bottom in columns. The chi-square value for Table V is 115.8, demonstrating that the overall distribution is significant (shows an association between the variables of person and voice) to $p < 0.001$ (p being the probability of error). As stated in the previous chapter, the Fisher Exact Test tests the significance of differences between two individual squares by providing a p-value; a p-value of less than 0.05 is considered significant, and the lower the p-value the greater the significance. This tests show no significant difference between the inputs in the top row since these are both zeros. The difference between the inputs in the bottom row is significant to $p= 0.004$ with the Fisher Exact test. The difference between the inputs in the left column is significant to $p=0.005$ with the Fisher Exact. Finally, the difference in the right column is significant to 0.0 with the Fisher Exact. To show that these results were not the result of the design decision to exclude non-passivizing verbs, I also calculated figures in which all verbs were included. The results of the Fisher Exact Test actually showed increased significance for all pairs of squares.

3.4.3 Results & Analysis: First vs. Second vs. Third Person

Table VII. Raw Data: Local Persons Split Into First and Second Person

Agent ↓ Patient→	First Person	Second Person	Third Person
First Person	A: 17 (17+ 0 top) P: 0 (0 + 0 top)	A: 146 (146 +0 top) P: 0 (0 +0 top)	A: 5205 (5182 +23 top) P: 0 (0 +0 top)
Second Person	A: 11 (11 + 0 top) P: 0 (0 + 0 top)	A: 5 (5 +0 top) P: 0 (0 +0 top)	A: 1041 (1035 +6 top) P: 0 (0 +0 top)
Third Person	A: 340 (340 + 0 top) P: 14 (14 + 0 top)	A: 132 (132 +0 top) P: 0 (0 +0 top)	A: 3110 (3107 +3 top) P: 39 (39 +0 top)

Table VIII. Rates of Passivization: Local Persons Split Into First and Second Person

Agent ↓ Patient→	First Person	Second Person	Third Person
First Person	0%	0%	0%
Second Person	0%	0%	0%
Third Person	4.0%	0%	1.2%

Table VIII. Rates of Passivization:
Local Persons Split Into First and Second Person

A ↓ P→	First Person	Second Person	Third Person
First Person	0%	0%	0%
Second Person	0%	0%	0%
Third Person	4.0%	0%	1.2%

While most of the numbers in the table are too small to provide significant results, the results which do rise to significance suggest the person hierarchy 1 > 2 > 3 in English. Inter-square comparisons in the first two columns provide no significant differences, but the data from the third column suggests that both first person and second person are ranked above third person. This is evidenced by significantly higher rates of passivization with the third-person patient in this column than the local-person patients. Comparing the top right square with the bottom right square gives p=0 on the Fisher Exact Test; this supports 1 > 3. Comparing the bottom right square with the one immediately above it gives p=0.000013 on the Fisher test; this supports 2 > 3. The data in the third row at first seem problematic; they suggest 1 > 3 > 2 because the rate of passivization should decline from left-to-right. However, examining significant differences does not support this. Comparison of the bottom left square with the bottom right square gives p=0.00039 using the Fisher test; this supports 1 > 3. Comparison of the bottom left square with the square immediately to the right gives p=0.01 using the Fisher test; this supports 1 > 2. However, comparing the bottom right square with the square immediately to its left gives an insignificant p= 0.199685 using the Fisher Test. Therefore there is no significant evidence for 3 > 2 (which would have been problematic). Overall, analysis of all the significant differences points to the person hierarchy 1 > 2 > 3 in spoken English.

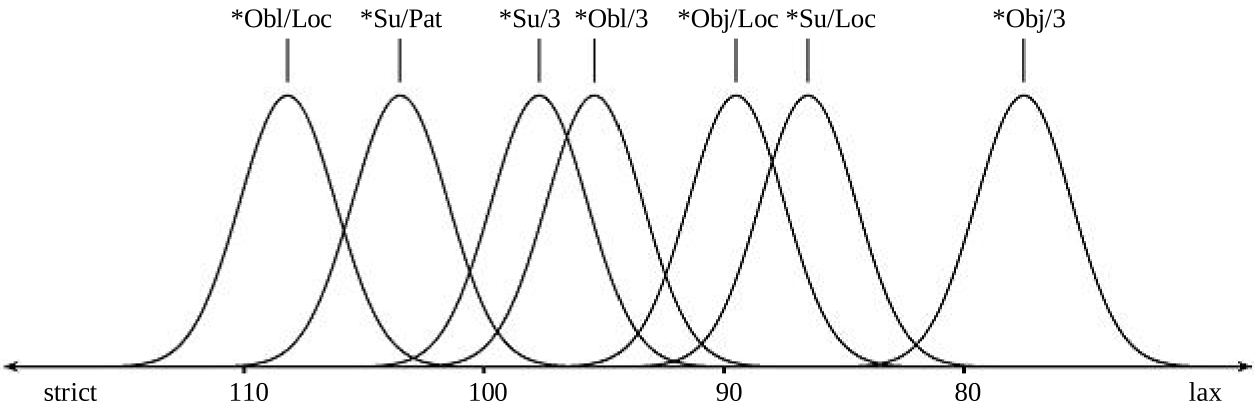
3.5 Stochastic OT Analysis

The constraints from the alignment of the person hierarchy with the grammatical relations hierarchy, together with the constraints on the association of semantic role with grammatical relations, were trained using the gradual learning algorithm on the (first set of) data obtained from the Switchboard Corpus. Subhierarchies were maintained. This produced the constraint rankings below.

Table IX. Constraint Rankings After Training

*Su/3	97.7
*Su/Local	86.5
*Oblique/Local	108.2
*Oblique/3	95.4
*Object/Local	89.5
*Object/3	77.5
*Su/Pat	103.5
*Su/Agt	96.5
*Obj/Agt	100.0
*Obj/Pat	96.5

These rankings are depicted graphically below. For simplicity, the constraints concerning semantic roles were omitted, except for the constraint *Su/Pat.



These constraint rankings produce the following percentages of passive. The figures in parentheses represent the original distribution:

Table XI. Rate of Passivization Produced By the Learned Constraint Rankings

Agent ↓ Patient→	Local Person	Third Person
Local Person	**0.0%** (0%)	**0.0%** (0%)
Third Person	**2.8%** (2.9%)	**1.1%** (1.2%)

As can be seen, the output distributions closely match the input distributions. The ranking of *Oblique/Local at 108.2 makes it the highest-ranked constraint; it is more than ten units above any constraint that would favor the passive when the agent is a local person (constraints favoring the passive would be *Obj/3 for a third person patient and *Obj/Local for a local person patient), accounting for the zero rate of passivization in the top row. Again, the ranking of *Oblique/Local over ten units above constraints favoring passive with local agents is problematic because passives with local person agents are grammatical in English; this highlights the need for topicality constraints and more training data for a better model.

Now consider the second row. With grammatical role constraints alone, the rate of passivization would be approximately 1.3%, due to the high ranking of *Su/Pat at 103.5 (the second-highest constraint) and lower-ranked constraints penalizing active. However, the high ranking of *Su/3 at 97.7 leads this to be elevated on the left (because *Su/3 penalizes active in this case). This is slightly tempered by *Obl/3 ranked at 95.4, which penalizes passives in the bottom left cell. The other constraints play less of a role. In the bottom right cell, the constraint *Su/3 has no influence since both active and passive violate it. Therefore, there is no constraint penalizing the active except for the grammatical role constraints (*Obj/Pat, *Su/Agt at 96.5) already mentioned; if it were not for these constraints the passive would occur at 0%. Again, this highlights the need for topicality constraints, since the 1.3% rate of passivization in the bottom right cell is presumably due not to an occasional dispreference for active, but to higher topicality of the third person patient in that case.

The high ranking of *Oblique/Local over *Obj/Local, puts English in line with most of the languages manifesting person-voice interactions discussed above; that is, with /local person patient + local person agent/, the passive is dispreferred. Also, while only a preliminary observation, it is interesting to note that *Su/Pat is ranked at 103.5 when trained on the above data from the Switchboard Corpus while ranked at 100.6 when trained on data from the Wall Street Journal Corpus – this could reflect the lower tendency to passivize in the spoken register. However, since the same constraint systems were not used in each case, this is uncertain.

3.6 Conclusion

In this chapter it has been demonstrated that the person hierarchy has effects on the frequency of passivization in English, and that these effects are consistent with the frequency-gradation predictions of constraints resulting from harmonic alignment of the person and grammatical relations hierarchies (as in Aissen 1999) in a Stochastic Optimality Theory framework. It has also been shown that these frequencies can be modeled within Stochastic Optimality Theory. Finally, examination of corpus data has faintly suggested that the cross-linguistically varying ranking of first and second persons in spoken English takes the form 1 > 2 > 3.

Appendix

Script Definitions of Different Types of Noun Phrases

```
$localPronouns="me|Me|us|Us|myself|Myself|yourself|Yourself|yourselves|Yourselves|ourselves|Ourselves|I|you|You|we|We";
$firstPronouns="me|Me|us|Us|myself|Myself|ourselves|Ourselves|I|we|We";
$secondPronouns="yourself|Yourself|yourselves|Yourselves|you|You";

$localNP=" !<<, (__>> (/^NP/ \$ CC \$ /^NP/)) <<, ($localPronouns > PRP)";
$firstNP=" !<<, (__>> (/^NP/ \$ CC \$ /^NP/)) <<, ($firstPronouns > PRP)";
$secondNP=" !<<, (__>> (/^NP/ \$ CC \$ /^NP/)) <<, ($secondPronouns > PRP)";
$thirdNP="!<<, (__ >> (/^NP/ \$ CC \$ /^NP/)) !< EX !< /NONE/ !<<, ($localPronouns > PRP)";
```

Verbs Judged to be Nonpassivizing in the Majority of Their Senses

afford
average
averages
averaging
balance
balanced
balancing
began
begin
blew
blow
blown
blows
care
charge
charged
charges
charging
chat
come
comes
commit
commits
committed
complain
contain
cost
costed
costing
costs
'd
darned
dating
die
escape
fit
get
gets
getting
go
goes
going
gone
got
gotten
graduated
graduating
had
Had
has
have
having
hemorrhaged
involve
involved
involves
jump
jumping
last
lasted
Leaned
live
lived
living
mind
missing
quit
ride
rides
riding
rode
rotted
rough
roughed
roughing
sat
sit
sitting
squeaked
stand
sub
subbed
suffices
talk
talked
talking
talks
've
vote
voted
waited
walk
walked
want
wanted
wanting
wants
weighed
weighs
went
work
worked
working
works

Verbs Excluded Due to Lack of Corresponding Active Form

born

Verbs Included

abandoned
absorb
absorbed
abuse
abused
accept
accepted
access
accused
accusing
add
added
address
addressed
adds
adjust
admire
admit
admitted
adopted
adored
advance
advertise
advertised
advise
age
ages
aggravate
aggravates
aimed
alleviate
alleviated
alleviates
allot
allow
allows
amaze
amazed
amazes
amended
annoys
anointed
answer
anticipate
appalls
applaud
apply
applying
appreciate
appreciated
approach
approved
ask
asked
asking
assign
assisting
astonished
astounded
ate
attached
attack
attempting
attend
attended
attending
attract
attribute
auctioned
audit
avoided
award
awarded
awarding
baby-sitting
back
backed
backlog
bake
baking
bank
banked
bankrupted
bar
bargained
barred
base
based
beat
beating
beats
beeped
befriends
beg
begging
believe
believed
believing
benefit
bet
bill
bit
bite
bites
bitten
blame
blaming
blasts
blesses
blinded
board
boil
bombarded
bored
borrow
borrowing
bother
bothered
bothers
bought
braved
break
breaking
breaks
bred
breed
breeding
bring
bringing
brings
broadcast
broadcasting
broadened
broke
broken
brought
brown
brush
budget
budgeted
budgeting
bug
build
building
builds
built
bundle
burned
burning
burns
burnt
bury
busting
buy
buying
buys
calculate
calibrated
call
called
calling
calls
calmed
calms
cancelled
carbureted
carried
carry
carrying
cashes
catch
catching
caught
cause
causes
causing
celebrated
centered
chained
change
changed
changes
changing
chased
check
checked
chewed
choose
chop
chopped
chose
chosen
chuck
circumventing
claiming
classify
clean
cleaned
cleaning
clip
close
closed
clued
coaches
collect
collected
collecting
combine
commend
compare
compared
comparing
completed
complicate
complicated
concerned
concerns
condemning
conditions
consider
considered
considering
contacted
contaminate
continue
contribute
contributed
contributing
control
controls
convert
converted
convict
convicted
convicting
convinced
cook
cooking
cooks
core
correlate
count
counted
cover
covered
covering
covers
crack
cracked
create
created
creates
criticizing
crocheted
crossed
crosses
crush
crushes
crushing
curbs
curtailed
cut
cuts
cutting
damage
damaged
deactivated
decide
decided
declare
declined
decreases
deduct
deducting
defeated
defeating
defeats
define
defray
delivered
demand
demonstrated
denied
dent
deny
denying
deplete
describes
deserted
deserve
designed
destroyed
destroys
determined
devastate
devastated
develop
developed
devise
devoted
devoured
diagnosed
dial
did
diminishing
dip
direct
directed
disables
disabused
disappoints
disconnect
disconnected
discontinued
discovered
discovers
discuss
discussed
dislike
dispelled
displayed
distributing
divided
dividing

divorced
divulging
divvy
do
does
doing
dominated
donate
donated
done
double
doubt
dowels
drafted
drafting
drag
dragged
drags
drank
drawing
dressed
drew
drink
drip
drive
driven
drives
driving
drop
dropped
drops
drove
dry
dug
dump
dumped
dumping
dumps
earn
earning
earns
eat
eaten
eating
eats
edit
educate
effects
elect
elected
elicited
eliminate
eliminated
eliminates
eliminating
emancipate
emit
empaneled
emphasized
emphasizing
empty
encompass
encourage
end
enforce
enforced
enjoy
enjoyed
enjoying
enjoys
enter
entered
entertained
envision
envy
equals
equated
escalated
established
evacuated
evoke
examines
exceed
excluded
excluding
excusing
execute
exonerated
expanding
expect
expected
expecting
expended
experience
experienced
explains
explored
expresses
expressing
extended
extending
extradited
faced
facing
failing
favor
fax
fear
feed
feeding
feeds
feel
feeling
felt
fertilize
fight
fighting
figure
figured
figuring
file
fill
filled
filling
fills
filmed
filter
find
finding
finds
finish
finished
finishing
fire
fired
fix
fixed
fixes
fixing
flew
flipping
float
folded
follow
followed
following
follows
forbid
force
foreclosed
forested
forget
forgot
formed
forward
forwarded
fought
foul
found
frame
framed
freaks
free
freeze
fried
frightened
frightens
froze
frustrated
frustrates
fry
fund
funded
gain
gained
gaining
gather
gathering
gave
generate
give
given
gives
giving
glued
goofed
gouging
grab
grabs
grasp
greets
grill
grips
grow
grown
grows
guarantee
guarantees
guard
guarded
guess
guessed
guessing
guided
hack
hammered
hand
handle
handled
handles
harm
harmed
hate
hated
hates
haul
heading
heads
hear
heard
hearing
heat
heated
held
help
helped
helping
helps
hire
hired
hires
hit
hits
hitting
hold
holding
honor
hooked
hooking
hosted
housesat
hugs
hung
hurt
hurting
identified
identify
ignore
ignored
imagine
imagined
impact
implies
impress
impressed
improve
incapacitate
include
included
includes
including
increase
increased
indicted
indoctrinated
induced
inducted
infatuated
influence
influenced
informed
inherited
injured
injures
inspect
installed
instituted
insulating
interest
interested
interpret
interrupted
interviewed
intrigued
introduced
introducing
invaded
invades
invading
inventing
invest
irks
iron
irritates
isolate
issued
jeopardizing
join
joined
joining
joins
judge
justify
keep
keeping
keeps
kept
kick
kicked
kidding
kill
killed
killing
kills
knew
knock
knocked
knocks
know
known
knows
label
lacked
laid
landscape
landscaped
lay
layer
laying
lead
leads
learn
learned
learning
lease
leasing

leave
leaves
leaving
led
left
lessened
let
level
licked
lie
lift
like
liked
liken
likes
liking
limit
limited
limits
lined
list
listed
lit
loaded
loads
loaned
lobby
located
lock
locked
look
loose
lose
loses
losing
lost
love
loved
loves
loving
lowered
lumped
made
mailed
maims
maintain
make
makes
making
manage
manages
managing
manipulated
manufacturing
marinate
marinates
mark
marked
market
married
marry
mashed
match
matches
mean
means
measured
meet
meeting
meets
melted
memorize
mention
mentioned
messed
met
microwave
minimize
misclassed
mishandle
mislocate
miss
missed
misses
misspelled
misuse
mix
mixed
models
moderated
mold
monitor
monitored
move
moved
moves
moving
mow
mowed
mulched
murdered
nail
nailed
name
need
needed
needlepointed
needs
neglected
netted
neuter
neutered
notice
noticed
noticing
notify
nursed
occupied
occupies
offend
offer
offered
offering
offers
open
opened
opens
operate
order
ordered
organize
organized
organizing
ostracized
outgrew
outgrown
outlawed
outlive
outselling
overheard
overheat
overlooking
overtaken
overwhelm
owe
own
owned
owns
pack
packaged
packed
paddle
paid
paint
painted
painting
papered
parole
paroled
pass
passed
patch
patrol
pay
paying
pays
peels
penalize
penalized
penalizes
perceive
perceived
performed
persuaded
pet
phase
pick
picked
picking
picks
piece
pinch
pirated
pitched
place
placed
plagued
plan
planned
planning
plant
planted
planting
plants
play
played
playing
plays
plead
please
pleased
pledged
point
pollute
poof
pooh-poohed
pop
popped
post
postpones
pour
pouring
practice
praised
preapproved
preceded
precedes
predicting
prefer
preferred
preferring
prepared
present
presented
presents
press
pressed
prevent
prevents
print
printed
prioritize
process
produced
produces
producing
program
prohibit
prohibited
promise
promised
promote
promoted
promoting
pronounce
propose
protect
protects
protesting
prove
provide
provides
providing
publicize
publicized
published
pull
pulled
pulling
pulls
pump
pumped
punch
punished
purchase
purchased
pursue
pursued
push
pushed
pushes
pushing
put
puts
putting
qualify
question
questioned
quilt
quilted
quote
quoted
rack
raise
raised
raises
raising
ran
raped
reached
reaching
read
reading
reads
realize
realized
rearrange
rearranging
receive
received
receiving
reclaimed
recognize
recognized
recommend
recommended
recommending
record
recorded
recycle
recycled
recycles
recycling
redone
reduce
re-evaluate
refer
referred
refined
refinished
reflected
reform
refund
refute
refuted
regain
registered
reimbursed
reinstated
reject
release
released
releasing
relinquished
remember
remembers

remind
reminded
reminds
remodeling
remove
removed
rendered
renewed
renovated
rent
rented
reopened
repacked
repainted
repairs
repealed
repeat
replaced
replacing
report
reported
represent
requested
require
requires
requiring
rereading
rescues
resembled
resisted
respect
respects
restored
restoring
retain
retains
retired
retitled
return
returned
returning
reverse
review
revived
revoked
rewired
rid
rig
righted
rinse
rip
ripped
risk
roast
rob
roll
rolled
romanticize
rotate
route
rub
ruined
ruining
run
running
runs
rush
rushed
rushes
sacrificed
sacrificing
said
save
saved
saves
saving
saw
say
saying
says
scan
scare
scared
scares
schooling
scold
scoot
score
scouting
scraped
scratch
scratched
screened
screening
screw
screwed
screws
sectioned
sectioning
see
seeing
seen
selected
sell
selling
send
sending
sends
sent
sentence
sentenced
sentencing
separate
separated
serve
served
serves
set
sewed
shade
shaken
shaped
share
shared
shaved
shed
shell
shifts
shoot
shooting
shorted
shorten
shot
shovel
show
showed
showing
shown
shows
shred
shut
sign
signed
singing
single
sings
skied
skip
skipped
slice
slipped
slow
smell
smelled
smoke
snatched
sneak
socked
softens
sold
solve
solved
sort
spank
spanked
spare
spared
spawn
speak
speaks
speeded
spell
spelled
spend
spending
spends
spent
spitted
split
splitting
sponsoring
spot
spray
sprayed
spread
spreading
sprinkle
spur
stack
staff
start
started
starting
starts
stash
stationed
steam
sterilize
stick
stiffen
stifle
stitch
stocked
stoned
stop
stopped
store
straightened
stranded
strangled
stress
stretched
stretching
strip
stripped
struck
stuck
studied
study
studying
subcontracted
subpoena
subtract
suck
sue
sued
suggested
suggesting
suing
summonsed
supplement
support
supported
surprise
surprised
surprises
surrounding
surrounds
suspect
swatted
swept
switch
switched
swung
synthesizing
tack
tackling
tailing
take
taken
takes
taking
tally
tape
taped
taping
tapped
target
targeted
taught
tax
taxed
taxes
teach
teaches
teaching
tear
tearing
teased
teasing
tell
telling
tells
tempered
tempted
term
terminate
terrifies
test
tested
testing
textured
thank
thanking
think
thinking
thinks
thought
thrashed
threw
throw
throwing
thrown
throws
thwarted
ticketed
ticks
tie
tied
till
time
tires
toast
told
tolerate
took
tore
torn
totaled
touch
touched
touches
tow
tows
trace
tracking
trade
traded
trained
training
transfer
transferred
trapped
treat
treated
tried
tries
trimmed
troubles
trust
try
trying
tuned
turn
turned
turning

turns
twirling
twirls
type
typed
types
understand
understood
undertook
undo
unfold
unload
update
upgraded
upheld
upset
use
used
uses
using
utilize
vary
vegetate
videotape
violate
visit
visited
visiting
visualize
warned
wash
waste
wasted
wasting
watch
watched
watches
watching
wear
wearing
wears
weave
wedded
weed
wet
win
winnowed
wipe
wish
won
wondering
word
wore
worries
wrap
wrapped
wreck
wrecked
write
writes
writing
written
wrote
yanked
zaps

Chapter 4

Conclusion

In what follows I consider possible objections to the models presented in the preceding chapters. I will argue, following Bresnan, Dingare and Manning (2001) that modeling effects of person and definiteness does require the person and definiteness constraints introduced previously. It will also be examined how the stochastic optimality theory model handles constraint "overlap" and how constraint overlap in turn restricts the typology of possible languages.

The discussions of the person and definiteness hierarchies in the previous chapters began with examination of how these hierarchies had been "grounded" by various researchers. It was shown that theories grounding the hierarchies in topicality and psycholinguistic accessibility were the most explanatory of their effects on the choice between active and passive. The question that immediately arises, then, and challenges the accounts given in the previous chapters, is why the observed frequencies of passivization should be dealt with by eighteen separate person and definiteness constraints rather than by a small number of constraints addressing topicality or accessibility. On such an account, the 1.1% rate of passivization with definite agents and indefinite patients would be explained by assuming that in exactly

1.1% of cases, the indefinite patient was more topical than the definite agent. In a discussion of English passives with indefinite subjects, Utsugi (1998) claims that this is precisely what occurs: that in those cases where passives occur with indefinite patients and definite agents, the indefinite patient is more topical than the definite agent. He provides the following example:

(1) There was a bomb threat at our college.
A student was questioned by the police. (Utsugi 1998 p.128)

In this example, the indefinite *a student* is more linked to the college mentioned in the previous sentence than the definite *the police*. This example demonstrates that indefinites can on occasion be more topical and accessible than definites.

The first problem that is raised by the above data is a problem for the idea that definites, indefinites, and so on are *syntactic coding devices* for topicality, the view of Givón (1983). If the same topicality that drove the choice of particular coding devices also drove the choice of passive, we should never see passives when the agent was higher on the definiteness hierarchy than the patient. This is because use of the passive implies that the patient is more topical than the agent, implying that the patient must be referred to with a more definite marker than the agent. The fact that passives do occur with patients that are lower on the definiteness hierarchy than agents demonstrates that the factors driving choice of referring expressions are not identical to the factors driving choice of passive. Related to this point is the point of Ariel (2000), who contends that local persons are not always highly accessible; they vary in accessibility, and are marked by zero pronouns (in languages that have them) when highly accessible, and by overt pronouns when less accessible. In English, zero subjects in declarative sentences are not available. This means that the local pronouns code a wider range of accessibility.

These facts support what is already clear, that constraints dealing with person and definiteness cannot fully explain the choice between active and passive. The question that remains is whether person and definiteness constraints play *any* role in explaining this choice, or whether the influence they seemed to have in the previous chapters was simply due to a correlation with other factors. One of the most forceful arguments that person and definiteness constraints play a role is languages in which certain definiteness-voice and person-voice configurations are ungrammatical (cf. sections 2.3 and 3.3.1). Bresnan, Dingare, and Manning (2001) review the argument that even in these languages, person-voice effects are driven by a grammaticization of topicality or newness constraints, eliminating the need for separate person constraints.

Consider the objections to this argument. Firstly, the simple fact that passives with local person agents do occur (cf. section 3.3.2) demonstrates that *in precisely those factors which drive passive*, whether it be newness or topicality or accessibility, local persons are not always the most given or topical or accessible (this is supported by Ariel's evidence that local persons vary in accessibility and by Givón (1994), who contends that local persons are statistically more likely, but not invariably, the topics of unmarked clauses). Unless we argue that the Lummi passive is driven entirely by newness while the English passive involves other factors, or that the inputs which occur in English simply do not occur in Lummi, we cannot maintain that the person-voice interactions in Lummi can be entirely accounted for with newness constraints. The second objection to accounting for person-voice effects with newness constraints is the grammaticization of animacy along with person in some of the languages that exhibit categorical person-voice effects, for example Chamorro[1] and Southern Tiwa. In these languages, it is ungrammatical to have inanimate subjects with animate objects. Clearly it would be impossible to argue that animates are *always* discourse-older than inanimates. Rather, they are only statistically more likely to be so (Cooreman 1983). Finally, as discussed in Bresnan, Dingare, and Manning (2001), the need for separate person constraints is motivated by the distinct treatment of first and second persons in some languages exhibiting person-voice interactions.

Thus, even if the grammaticization of person and animacy is driven by topicality, separate person and definiteness constraints are necessary for those languages in which categorical person-voice and definiteness-voice interactions occur. Yet this does not necessarily motivate them in English. Consider, however, what the grammaticization of person, animacy or definiteness implies. Presumably, in the languages that have a prohibition on local-person agents in *by*-phrases, the statistical tendency for local-person agents to be subjects hardened into a formal rule. That is, even in those cases where first-persons were relatively inaccessible, speakers began to realize them as subjects "out of habit". One reason to have person and definiteness constraints in English is that the same phenomenon could be occurring with English speakers: even when indefinite patients are more topical than definite agents, speakers may rule out passive due to the rarity of indefinite subjects and definite *by*-phrases. This amounts to saying that the speaker's past language experience influences the sentences he is likely to produce (the data-oriented-

[1] See Chung (1998) for arguments that person-voice effects in Chamorro are not categorical, but only statistical.

parsing model reviewed in Bod and Scha (1995)). Supporting the application of the language-experience model to notions such as animacy is evidence cited by Siewierska (1984) that sentences with definite and human subjects are more easily produced and understood than other types of subjects.

A similar story can be told for sympathetic effects across other hierarchies. It has often been noted that hierarchies overlap – if one posits a separate animacy hierarchy, thematic hierarchy, definiteness hierarchy, person hierarchy, and topicality hierarchy, one faces the problem that first persons are always topical and animate, agents tend to be animate and definite, patients tend to be inanimate, animates tend to be topical, and so on. While it has been argued that certain of the hierarchies drive the others (that grammaticization of person is due to grammaticization of newness, and that grammaticization of animacy is due to the topicality of animates), I have argued here that one needs constraints addressing them separately. Stochastic OT provides an elegant way to model how the separate constraints from different hierarchies influence one another.

Consider the tableau below. In this tableau, the topicality constraint *Su/Non-Topical (penalizing subjects which are non-topical in the discourse) and the constraint *Oblique/Local are exactly parallel and penalize passivization. This represents the most frequent situation when the input contains local person agents: the local person agent is also topical. The fact that the two constraints have exactly the same markings means that during training in the stochastic OT gradual learning algorithm, they will be treated the same (that is, demoted or promoted by the same amount). Since this input will be realized as active, both constraints will be pulled up above *Su/Local. Thus, even if the active output is in some sense driven by topicality, the person constraint will get pushed up as well. Since the topicality constraint will play a role with inputs in which the person constraint will not play a role, the topicality constraint will emerge with a different ranking. Nevertheless, if the input below is frequent, the person constraint will end up close enough to the topicality constraint to drive the choice of active even when the local person agent is non-topical.

Tableau I.

/Local Person Topical Agent- Third Person Non-Topical Patient/	*OBLIQUE/LOCAL	*SU/NON-TOPICAL	*SU/LOCAL
☞ [Active]			*
*[Passive]	*!	*	

That is, because the local person agent is rarely realized as an oblique, the speaker will disprefer passive even when the local person agent is non-topical.

This “sympathetic” behavior of constraints restricts the typology of possible languages. Rather than reranking constraints from different hierarchies in every possible way, the result of training with the gradual learning algorithm is that certain constraints will move together. This predicts that languages which realize agents as subjects may also tend to realize animates and definites as subjects even when they are not agents. Similarly, languages which realize topical elements as subjects may occasionally realize animates and local persons as subjects even when they are not topical. Thus, a language like English could not selectively show effects of the topicality hierarchy and simultaneously not show effects of the person hierarchy, and it is predicted that languages which show effects of the person hierarchy fall back on topicality to choose between active and passive when both inputs are third person. Most interestingly, this restriction in typology is due to properties of the input – the fact that agents and local persons are usually definite, animate, and topical.

In the previous chapters I have presented evidence that prominence hierarchies affect frequencies of passivization in English and demonstrated how stochastic optimality theory can be used to model these effects. Here I have defended against certain objections to the models and have also explored how stochastic optimality theory can be used to model “constraint overlap” effects.

Bibliography

Aissen, Judith. 1999. Markedness and subject choice in optimality theory. *Natural Language and Linguistic Theory*. 17: 673-711.

Aissen, Judith. 2000. Differential object marking: iconicity vs. economy [DRAFT]

Allen, B. J. and D. G. Frantz. 1978. Verb agreement in Southern Tiwa. *BLS* 4: 11-17.

Ariel, Mira. 1990. *Accessing Noun-Phrase Antecedents*. London: Routledge.

Ariel, Mira. 1996. Referring expressions and the +/- coreference distinction. In *Reference and Referent Accessibility*, ed. by Thorstein Fretheim and Jeanette Gundel. John Benjamins.

Ariel, Mira. 2000. The development of person agreement markers: from pronouns to higher accessibility markers. In *Usage-Based Models of Language*, ed. by Suzanne Kemmer and Michael Barlow, 197-260.

Birner, Betty J. and Gregory Ward. 1998. *Information Status and Noncanonical Word Order in English*. Amsterdam: John Benjamins.

Bod, Rens and Remko Scha. 1997. Data-oriented language processing: an overview.

Boersma, Paul and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*. 32:45-86.

Boersma, Paul and David Weenink. 2000. Praat computer program. On-line, Institute of Phonetic Sciences, University of Amsterdam: http://www.fon.hum.uva.nl/praat/.

Bresnan, Joan, Shipra Dingare, and Christopher Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG 01 Conference, The University of Hong Kong*, ed. by Miriam Butt and Tracy Holloway King. On-line, CSLI Publications: http://csli-publications.stanford.edu/.

Chung, Sandra. 1998. *The Design of Agreement: Evidence From Chamorro*. Chicago: University of Chicago Press.

Cooreman, Ann. 1983. Topic continuity and the voicing system of an ergative language: Chamorro. In Givón (ed.), 1983, 425-89.

Cooreman, Ann. 1987. *Transitivity and Discourse Continuity in Chamorro Narratives*. Berlin: Mouton de Gruyter.

DeLancey, Scott. 1981. An interpretation of split ergativity and related patterns. *Language* 57: 626-657.

Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67:547-619.

Dixon, R.M.W. 1979. Ergativity. *Language*. 55:59-138.

Estival, Dominique and John Myhill. 1988. Formal and functional aspects of the development from passive to ergative systems. In *Passive and Voice*, ed. by Masayoshi Shibatani, 441-91. Amsterdam: Benjamins.

Forrest, Linda B. 1994. The de-transitive clauses in Bella Coola: Passive vs. inverse. In *Voice and Inversion,* ed. by Talmy Givón, 147-68. Amsterdam: Benjamins.

Francis, H., M. Gregory and L. Michaelis. 1999. Are lexical subjects deviant? In CLS-99. University of Chicago.

Givón, Talmy. 1976. Topic, pronoun, and grammatical agreement. In *Topic and Subject*, ed. by Charles Li, 149-88. New York: Academic Press.

Givón, Talmy. 1979. *On Understanding Grammar*. New York: Academic Press.

Givón, Talmy. 1983. Topic continuity in discourse: an introduction. In Givón (ed.), 1983, 1-42.

Givón, Talmy. (ed.) 1983. *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. Amsterdam: Benjamins.

Givón, Talmy. 1994. The pragmatics of de-transitive voice: Functional and Typological aspects of inversion. In Givon (ed.), 1994, 3-44.

Givón, Talmy. (ed.) 1994. *Voice and Inversion*. Amsterdam: Benjamins.

Greenberg, Joseph H. 1966. *Language Universals: With Special Reference to Feature Hierarchies*. The Hague: Mouton.

Greenbaum, Sidney. 1980. Syntactic frequency and acceptability. In *Evidence and Argumentation in Linguistics*, ed. by Thomas A. Perry. Walter de Gruyter Berlin; New York

Jelinek, Eloise and Richard Demers. 1983. Agent hierarchy and voice in some Coast Salish languages. *International Journal of American Linguistics*. 49: 167-85.

Kato, Kazuo. 1979. Empathy and passive resistance. *Linguistic Inquiry*. 10: 149-152.

Klokeid, Terry J. 1978. Surface structure constraints and Nitinaht enclitics. *Linguistic Structures of Native Canada*, ed. by Eung-Do Cook and Jonothan Kaye, 157-76. Vancouver: University of British Columbia Press.

Kroskrity, P.V. 1985. A holistic understanding of Arizona Tewa passives. *Language* 61: 306-328.

Kuno , Susumu. 1987. *Functional Syntax: Anaphora, Discourse, and Empathy*. Chicago: University of Chicago Press.

Kuno, Susumu, and Etsuko Kaburaki. 1977. Empathy and syntax. *Linguistic Inquiry* 8: 627-72.

Manning, Christopher. 1996. *Ergativity*: *Argument Structure and Grammatical Relations*. Stanford: CSLI Publications.

Marcus, Mitchell, Beatrice Santorini, and May Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19:313-330.

McFarland. 1978. Definite objects and subject selection in Philippine languages. *Studies in Philippine Linguistics* 2: 139-82

Prince, Alan and Paul Smolensky. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*. RuCCS Technical Report No.2. Piscatway, NJ: Rutgers University Center for Cognitive Science.

Prince, Alan and Paul Smolensky. 1997. Optimality: from neural networks to universal grammar. *Science* 275: 1604-10.

Ransom, Evelyn. 1979. Definiteness and animacy constraints on passive and double-object constructions in English. *Glossa* 13:215-40.

Seoane Posse, Elena. 2000. The passive as an object foregrounding device in Early Modern English. In *Generative Theory and Corpus Studies*: *A Dialogue from 10 ICEHL,* ed. by Ricardo Bermudez-Otero, David Denison, Richard M. Hogg, C. B. McCully, 211-32. Berlin: Mouton de Gruyter.

Siewierska, Anna. 1984. *The Passive: A Comparative Linguistic Analysis*. London: Croom Helm.

Silverstein, Michael. 1986. Hierarchy of features and ergativity. In *Features and Projections*, ed. by Pieter Muysken and Henk van Riemsdijk, 163-232. Dordrect – Holland/Riverton: Foris Publications.

Svartvik, Jan. 1966. *On Voice in the English Verb*. The Hague: Mouton.

Tesar, Bruce and Paul Smolensky. 1996. *Learnability in Optimality Theory (short version)*. Report: JHU-CogSci-96-2. Baltimore, MD: Johns Hopkins University Department of Cognitive Science.

Utsugi, Aiko. 1998. A context-based account of English passives with indefinite subjects. In *Function and Structure*, ed. by Akio Kamio and Ken-ichiTakami, 123-36.

Wierzbicka, Anna. Case marking and human nature. *Australian Journal of Linguistics*. 1: 43-80.

Whistler, Kenneth W. 1985. Focus, perspective, and inverse person marking in Nootkan. In *Grammar Inside and Outside the Clause*: *Some Approaches to Theory from the Field*, ed. by Johanna Nichols and Anthony C. Woodbury, 227-65. Cambridge: Cambridge University Press.

Winter, Werner. 1971. Formal frequency and linguistic change: Some preliminary comments. *Folia Linguistica*: *Acta Societatis Linguisticae Europaeae*. 5: 55-61.

Zaharlick, Amy. 1982. Tanoan studies: passive sentences in Picurís. *Ohio State University Working Papers in Linguistics* 26: 34-48.